Hierarchical Bayesian ANOVA for Multivariate and Spatial Data with Application to Climate Projections

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Steven Richard Geinitz

aus den

Vereinigten Staaten von Amerika

Promotionskomitee

Prof. Dr. Reinhard Furrer (Vorsitz)
Prof. Dr. Leonhard Held

Zürich, 2013

Preface

The path taken to arrive at this point has not been easy, nor direct. It simply would not have been possible without the help of many people along the way that have added their own unique form of guidance, support, and motivation.

Thank you to Reinhard Furrer for offering me the opportunity to be the first doctoral student to begin working with him, for his encouragement and patience, and for providing the chance to move with him from the Colorado School of Mines to the University of Zurich. Studying in Switzerland has been one of the most enlightening experiences of my life.

A special mention must also be made of two individuals who had a brief, albeit profound influence. Thank you to Lee Davis, for your support and belief in me. Thank you to Bernard Gillette, for the indoctrination into two worlds, climbing and mathematics, both of which I cherish deeply. It is not possible to overstate the impact from them both.

Thank you to all of the friends and colleagues that I have met at the University, and in Zurich. It has been terrific getting to know and work with such a diverse group of individuals.

Thank you to all of my family. To my mom, dad, and my brother, you helped make me into the kind of person that would want to start this kind of program. To Melissa, thank you for your love and support, and for reminding me to, "stick to the job." To Eddie, thank you for sitting patiently beside me when things were going smoothly, and for suggesting a walk outside, when a break was needed.

Finally, thank you to the URPP for supporting this position within the Mathematics Institute. I have benefited greatly from this program, and I hope that others in the future will be able to as well.

Zurich, March 2013

Steven Geinitz

Zusammenfassung

Angewandte Statistik als eigenständige Fachrichtung wurde vor circa einem Jahrhundert anerkannt. Ein grosser Teil dieser Anerkennung kann der Analyse der Varianz zugeschrieben werden. Diese Methode hatte eine entscheidende Bedeutung für einen Grossteil des wissenschaftlichen und industriellen Fortschritts im 20. Jahrhundert. Da sich jedoch die Komplexität der Experimente und Daten während dieser Zeit erhöht hat, ist auch die Schwierigkeit bei der Durchführung einer Analyse der Varianz gewachsen. Oft ist es schwierig zu verstehen, wie statistische Parameter der realen Welt Mengen entsprechen; wie statistische Methoden miteinander in Beziehung stehen; zwischen dem Praxisbezug und der statistischen Signifikanz einer Analyse zu unterscheiden; und den Rechenaufwand einer gewählten Methode zu erkennen. Diese Arbeit leistet einen Beitrag auf dem Gebiet der angewandten Statistik durch die Festlegung einer Methode, die sich diese Herausforderungen annimmt. Diese Methode, die rechnerisch effizient ist, klare und intuitive Schlussfolgerungen liefert, wird dann auf simulierte Klimaprojektionen angewendet, um herauszufinden, welche Faktoren zur allgemeinen Variabilität der Daten beitragen.

Die Arbeit beginnt mit der Darstellung der Analyse der Varianz in einer allgemeineren Form als es üblich gemacht wird, in einer die die Verbindung zwischen *fixed* und *random* Effekts innerhalb eines einzigen Verfahrens erleichtert. Um die Ergebnisse zu presentieren werden neuartige numerische und grafische Darstellungen der Varianz-Parameter entwickelt, die die praktische Relevanz sowie die statistische Signifikanz aufzeigen. Es werden auch neuartige Darstellungen für allgemeine multivariate Fälle vorgeschlagen. Bestehende Methoden, diese Ziele zu erreichen, verlassen sich in der Regel auf Markov chain Monte Carlo Methoden. Im Gegensatz dazu, wird in dieser Arbeit eine Klasse von konjugierten Verteilungen entwickelt, die eine Zerlegung der Bayesian posterior Verteilung ohne MCMC erlaubt. Für Fälle, in denen räumlich-geostatistische Modelle vorliegen, werden Approximationsverfahren entwickelt, die den Rechenaufwand noch weiter verkleinern. Diese Approximationsverfahren werden genau untersucht und es wird gezeigt, dass sie genaue Darstellungen der ursprünglichen Daten liefern.

Der Höhepunkt dieser Bemühungen ist die räumliche und statistische Analyse einer Reihe von simulierten Klimaprojektionen. Eine einzelne simulierte Klimaprojektion wird mit einem deutlichen mathematischen Klimamodell und einer angenommenen Rate von zukünftigen Treibhausgasemissionen durchgeführt. Die Quellen der Variabilität vom Klimamodell und von der Emissionsszenario werden untersucht, sowie die Interaktion der beiden. Alle davon werden als Konstante Effekt untersucht, sowie als zeitliche Effekt ueber die dekadische Zeitspanne. Das Ausmass der Variabilität für jede Quelle wird dann geschätzt, zusammen mit der Unsicherheit dieser Schätzungen. Obwohl das Verfahren originell ist, stimmen die Ergebnisse mit denen in früheren Peer-Review-Arbeiten überein.

Abstract

Applied statistics as an independent field of study can be said to have been recognized as such roughly a century ago. A large part of this recognition can be attributed to analysis of variance, a method which has been of paramount importance for much of the scientific and industrial progress throughout the 20th century. However, as the complexity of experiments and data has increased during this time, so has the difficulty in carrying out an analysis of variance. Often it is difficult to understand how statistical parameters correspond to real-world quantities; to relate statistical methods to one another; to distinguish between the practical relevance and statistical significance of an analysis; and to recognize the computational demands of a chosen method. The work herein contributes to the field of applied statistics by identifying a methodology that addresses these challenges. This method is then applied to large-scale simulated climate projections in a way that is computationally efficient and that provides clear and intuitive conclusions on what explanatory factors contribute to the overall variability of the data.

The work begins with viewing analysis of variance as a more general procedure than it is traditionally considered. By placing it into a more broadly defined hierarchical Bayesian context, the ability to draw connections between fixed and random effects within a single procedure is facilitated. To present results, novel numerical and graphical summaries that communicate practical relevance, as well as the statistical significance, of variance parameters are then developed. Novel summaries for general multivariate cases are also proposed. Existing methods that achieve these goals typically rely on Markov chain Monte Carlo methods. Whereas here a class of conjugate distributions for variance parameters is derived, which allow for a factorization of the posterior distribution and which obviate the need for MCMC methods, thus reducing computational demands. For cases involving spatial geostatistical models, approximation methods that further alleviate computational demands are also developed. These approximations are closely examined, and shown to produce accurate representations of the original quantities.

The culmination of these efforts is in the spatial statistical analysis of a set of simulated climate projections. A single simulated climate projection is carried out using a distinct mathematical climate model, and an assumed rate of future greenhouse gas emissions. Thus, the sources of variability examined in the data are the climate model and the emissions scenario, their interaction, both as a constant as well as over decadal time increments. The magnitude of the variability for each source is then estimated, together with the uncertainty of these estimates. While the method is original, the results agree with those found in previous peer-reviewed works.

Thesis Outline

Introduction

Paper I: **Conjugate distributions in hierarchical Bayesian ANOVA for computational efficiency and assessments of both practical and statistical significance**
Geinitz, S. and Furrer, R.
Technical report, University of Zurich.

Paper II: **Multivariate analysis of global climate projections via rank deficient Bayesian ANOVA**
Geinitz, S., Furrer, R. and Sain, S. R.
Paper to be submitted.

Paper III: **Bayesian analysis of variance for relative assessment of general circulation model effect**
Geinitz, S., Furrer, R. and Sain, S. R.
Paper under revision for resubmission to the *International Journal of Climatology*.

Paper IV: **Assessing variance components of general circulation model output fields**
Furrer, R., Geinitz, S. and Sain, S. R.
Paper published in *Environmetrics* (2012), 23, pp. 440–450.

Introduction

It has long been recognized that the earth's climate is constantly evolving, that it is itself a living entity (Peixoto and Oort, 1992). Thus, there has long been a desire to understand and forecast its behavior. Attempts to predict climate are, however, not only scientific, nor economic endeavors, but rather a necessary undertaking that must be carried out in order to assess the possible ecological, political, and social impacts in the coming decades. Furthermore, the need to quantify the extent to which anthropogenic effects, i.e. effects due to human activity, drive climate behavior is becoming increasingly more urgent.

The term climate refers to the physical state of the earth's climatic system averaged over a specified amount of time. The climatic system is made up of several components, including the atmosphere (strata of gases), hydrosphere (water masses), cryosphere (ice and snow cover), lithosphere (crust and tectonic plates), and the biosphere (flora and fauna). The atmosphere, being the most variable of these, receives the greatest amount of focus when examining observable weather patterns (Daley, 1993). Precisely speaking, the term weather refers to the instantaneous state of the atmosphere. While many different properties of the state may be considered, it is typically classified on the basis of temperature and precipitation (Peixoto and Oort, 1992). As a result, climate herein is regarded as averaged quantities of temperature and precipitation over time scales on the order of one to two decades.

Because climate is a direct consequence of weather, any attempt to understand the former must begin with the latter. A result of paramount importance in numerical weather prediction (NWP) was to focus on the problem of integrating the equations of motion of the atmosphere (Bjerknes, 1904). Although initial attempts using this approach were unsuccessful, NWP has been further refined by atmospheric scientists and facilitated by advances in computation. However, the variability of these forecasts still must be recognized. For example, even with accurate reporting of the initial conditions, which are used as input for NWP, forecasts can become more or less random after a short amount of time. Although this time has increased over the years, together with increases in model improvements and with computational capabilities, it is still on the order of days. The high sensitivity from initial values is due to the nonlinearity of the equations involved and has serious implications for forecasts further into the future (Lorenz,

1972). Thus, understanding the uncertainty in atmospheric modeling is not only examined by climate scientists, but has begun to be investigated more closely by statisticians as well. See Kalnay (2002) for a more thorough review of the early developments of this field.

The systems of partial differential equations based on the thermodynamics and fluid mechanics on the globe, together with their coupling and numerical methods, form what is referred to as an atmosphere-ocean general circulation model (AOGCM or GCM). The full system of equations is discretized in space and in time, then carried out on some of the highest performance computers available (Nychka, 2000). One such GCM is the Community Climate Model (Kiehl *et al.*, 1998), which in its latest version used a spatial grid consisting of roughly $8,000$ points that are approximately $2.8^\circ \times 2.8^\circ$ with 20 minute time increments.

The scientific effort is enormous and the push to generate accurate models is not only academic in nature. Because of the potentially devastating effects of future climate changes on societies and economies worldwide, there is also a large effort by governments and international agencies to assess and discuss the results of GCM output. The culmination of this effort is seen in the Intergovernmental Panel on Climate Change (IPCC) report that is generated roughly every five to seven years. The first report was completed in 1990 (Houghton *et al.*, 1990). The and most recently completed was in 2007 and is referred to as the Fourth Assessment Report (AR4). Due to the enormous amount of information presented in the report it is is divided into three working group reports. The first, Working Group I, is titled *The Physical Science Basis* (Solomon *et al.*, 2007) and makes use of the current climate science knowledge, and analyzes the simulation results from a scientific point of view. Working Group II, *Impacts, Adaptation, and Vulnerability*, extends the results of the former by investigating how changes in climate could result in changes to the earth in other ways, e.g. rising sea levels, ecosystems changes, etc. Working Group III, *Mitigation of Climate Change*, addresses changes to social and economic policies that can be recommended in order to alleviate negative effects of a changing climate. Additionally, there is a synthesis report that summarizes the previous three. The next IPCC report (AR5) will be available soon. As with the previous reports, AR5 will be released in stages, with the first working group report scheduled for September 2013, and final synthesis report scheduled for October 2014.

The report most relevant for the work herein is the AR4 Working Group I Report. In this report there are 17 climate research institutions represented. Each of these has implemented at least one GCM for a total of 23. A significant portion of all IPCC reports are devoted to the anthropogenic effects on future climate. That is, how different rates of globalization and industrialization in developing countries affect the composition of the atmosphere, and thus

affect climate. As a result, GCMs are typically run under several different possible emission pathways for varying degrees of human activity in various respects, e.g. population growth, burning of fossil fuels, etc. These have been classified into emissions scenarios with each GCM carrying out climate projections under each given scenario (Nakićenović and Swart, 2000). A vital step in being able to accurately interpret the resulting climate projection data is to be able to distinguish the differences and variation among the GCMs and the scenarios.

An instance of such climate projection data is temperature (° Celsius) over the boreal summer months of June, July, and August. An individual datum, or so-called observation, is expressed as a vector existing in $\mathbb{R}^d$ with the d elements representing locations on a spatial grid over the earth. Note that the term *observation* may be misleading from a climate science point of view, as it does not refer to meteorological data that has been observed or measured. Rather, from the statistical point of view that is used herein, they refer to the data that are modeled statistically, i.e. simulated climate projections are the observations. Individual observations are denoted as $\mathbf{Y}_{ijt}$, where i denotes a GCM, j an emissions scenario, t a decade over which temperature has been averaged, and $d = 8192$ corresponding to a 2.8° spatial resolution. To illustrate the data, several summary statistics are calculated using a set of observations consisting of 15 GCMs, 3 emissions scenarios, and 9 decadal periods from 2010 to 2090. These are shown in Figure 1.

The spatial scale of the climate data generated by a GCM is on such a coarse grid that it is not currently practically possible to predict local climate behavior. Rather, the data from a GCM is sometimes used as input (boundary conditions) to a regional climate model (RCM), i.e. an atmospheric model on a finer resolution over a smaller geographical region. Similar to GCMs, there are many different implementations of RCMs created by different institutions. The two factors, GCM and RCM, then yield a crossed design that requires some understanding on which of the factors is responsible for more variation in the final output. Kaufman and Sain (2010) carry out a two-way spatial analysis of variance to assess variability from each of these sources. In the analysis they utilize finite-population and superpopulation variances with respect to spatial fields. Considering both finite and superpopulation variance parameters within a single statistical model is not new, but is more accessible with the use of a multilevel statistical model, or more generally, a hierarchical Bayesian model. Gelman (2005) illustrates this clearly and promotes the use of this type of analysis of variance model. These two references have been largely influential on the work herein.

It is necessary to mention that the term *model* is used heavily throughout the thesis, but in two quite different contexts. While both are considered a type of mathematical model, they are utilized differently herein. A *climate model*, which refers to a system of equations representing the

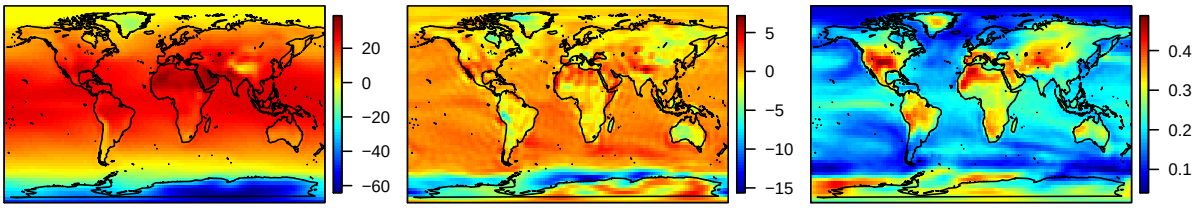

Figure 1: For the set of temperature data (°C) $\mathbf{Y}_{ijt}$, with i denoting a GCM, j an emissions scenario, and t a decadal average for periods 2010–2020, . . . , 2090–2100, summary statistics are calculated. The average over all GCMs, scenarios, and periods (left); a contrast of one GCM, the Canadian climate model CCCMA-CGCM3.1, (middle); and a contrast of one scenario, a moderate globalization scenario A1B, (right), show the different magnitudes across these factors, as well as how temperature varies spatially according to latitude, land cover, etc. In total, 15 GCMs, 3 emissions scenarios, and 9 decadal periods are represented.

thermodynamics and fluid mechanics on the earth, is what generates the data that is analyzed. A *statistical model* is what does the analysis, and which represents the observations in a manner that allows influential factors in the observations to be identified. When it is unclear from the context how the term model is being used, a distinction will be made.

The remainder of the introduction is structured as follows. A brief overview of hierarchical Bayesian analysis of variance as it is applied herein is given in Section 1. Additionally in this section, the novel technique of using constrained factor level prior distributions to obtain convenient factorized forms of the full joint posterior is alluded to. This is then covered in more detail in Section 2, as it is then utilized by the manuscripts to follow. In Section 3 various approaches of quantifying and comparing variability from multivariate sources of variation are covered. This is further discussed in Geinitz *et al.* (2012b). In Section 4 higher dimensional multivariate and spatial data is discussed. This includes spatial covariance matrices and approximations used to facilitate computation when working with large spatial fields, as is done in Geinitz *et al.* (2012a) and Furrer *et al.* (2012).

1 Analysis of Variance

In this section the standard and widely used method of analysis of variance (ANOVA) is briefly summarized. An overview of ANOVA in a hierarchical Bayesian framework is then given, along with some of its advantages over the classical approach.

1.1 Background of ANOVA

Analysis of variance is generally regarded as a technique to decompose data observations into a set of quadratic sums, according to the factors of interest in the statistical model (Fisher, 1925). For the one-way model with factor α and with the same number of replicates for each level of the factor, observations are denoted as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad i = 1, \ldots, n_I, \quad j = 1, \ldots, n_J, \tag{1.1}$$

where estimates of the parameters $\mu, \alpha_1, \ldots, \alpha_{n_I}$ are made by minimizing the quantity

$$\text{SSE} = \sum_{i,j} (Y_{ij} - \mu - \alpha_i)^2. \tag{1.2}$$

The solution is, however, not unique. Rather, constraints must be employed in order to identify a distinct solution. One such constraint commonly used on the estimates $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_{n_I}$ is that their sum is zero, i.e. $\sum_i \widehat{\alpha}_i = 0$. The resulting estimates are then $\widehat{\mu} = \overline{Y}_{..}$, which denotes the mean over all $n = n_I n_J$ observations, i.e. $\overline{Y}_{..} = \frac{1}{n} \sum_{ij} Y_{ij}$, and $\widehat{\alpha}_i = \overline{Y}_{i.} - \overline{Y}_{..}$, where $\overline{Y}_{i.} = \frac{1}{n_J} \sum_j Y_{ij}$. The sums of squares decomposition is

$$\sum_{i,j} Y_{ij}^2 = n_I n_J \overline{Y}_{..} + n_J \sum_i (\overline{Y}_{i.} - \overline{Y}_{..})^2 + \sum_{i,j} (Y_{ij} - \overline{Y}_{i.})^2, \tag{1.3}$$

which is, more precisely, referred to as the total sums of squares. What is commonly referred to simply as the sums of squares decomposition uses the mean adjusted quantities

$$\underbrace{\sum_{i,j} (Y_{ij} - \overline{Y}_{..})^2}_{\text{SST}} = n_J \underbrace{\sum_{i} (\overline{Y}_{i.} - \overline{Y}_{..})^2}_{\text{SSA}} + \underbrace{\sum_{i,j} (Y_{ij} - \overline{Y}_{i.})^2}_{\text{SSE}}. \tag{1.4}$$

The statistical model specification given by (1.1) generally assumes that $\epsilon_{ij} \sim \text{N}(0, \sigma_\epsilon^2)$, are independent of one another, and that α_i are fixed, but unknown constants. Hence, this is also referred to as a *fixed* effects model. The objective of the standard ANOVA procedure is then to check if there is a statistically significant difference among the α_i terms. Since SSE is proportional to a χ^2 distribution, and SSA is an offset of this, this may be tested with the hypothesis test, $H_0 \colon \alpha_1 = \cdots = \alpha_{n_I} = 0$. The test is carried out using the standard ANOVA table, and is shown in Table 1.1.

The probability given in the table is the p-value corresponding to H_0. The hypothesis will be rejected for large values of $F = \frac{\text{MSA}}{\text{MSE}}$. The hypothesis will not be rejected as F approaches unity, since the expected value of F is precisely equal to 1 when H_0 is true (Table 1.2). This feature of ANOVA is quite useful, as it allows for the testing of nested models, e.g. rejection of the null

Table 1.1: One-way analysis of variance summary.

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
α	$n_I - 1$	SSA	MSA $= \frac{SSA}{n_I - 1}$	$F = \frac{\text{MSA}}{\text{MSE}}$	$\Pr(F_{n_I-1, n-n_I} > F)$
ϵ	$n - n_I$	SSE	MSE $= \frac{SSE}{n-n_I}$		

hypothesis corresponding to the factor β for model $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$ would be evidence for its use over the use of the model given by (1.1).

In another scenario the terms α_i may be thought of as originating from a larger population, and an estimate of the uncertainty within the population is of interest. The statistical model uses again (1.1) and the same assumption on ϵ_{ij} terms, with the added assumption that $\alpha_i \sim \text{N}(0, \sigma_\alpha^2)$, and are independent of each other and of the ϵ_{ij}. This is generally referred to as a random effects model. The hypothesis test related to the previous one is $H_0 \colon \sigma_\alpha^2 = 0$. The test is carried out using the same F test statistic as before, since the expectation of F is again 1 when H_0 is true. Additional factors may be added in the statistical model, e.g. $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$, and

Table 1.2: Expectation of mean squares MSA and MSE.

Source	Df	Mean Sq	Fixed	Random
α	$n_I - 1$	MSA	$\sigma_\epsilon^2 + \frac{n_J}{n_I - 1} \sum_i \alpha_i^2$	$\sigma_\epsilon^2 + n_J \sigma_\alpha^2$
ϵ	$n - n_I$	MSE	σ_ϵ^2	σ_ϵ^2

follow the same procedure. Although it is one of the most recognized and relied upon methods in statistics, carrying out an ANOVA analysis is not always straightforward. Two issues are now discussed.

Practical and Statistical Significance

One of the most important features that classical ANOVA does not address is the ability to assess practical significance of a factor. This can be seen either as a trivial technical consequence of the statistical power that is inherent in standard ANOVA, or as an ideological fault of statisticians in continuing to promote methods that behave in this manner.

To examine how statistical significance is leveraged by sample sizes consider the model of (1.1), where $\text{Var}(Y_{ij}) = \sigma_\alpha^2 + \sigma_\epsilon^2$. Assume that $\sigma_\alpha^2 = \Delta \sigma_\epsilon^2$ so that the expectation of $F = \frac{\text{MSA}}{\text{MSE}}$ is $\mathbb{E}[F] = \frac{\nu_2}{\nu_2 - 2}(1 + \Delta n_J)$, where $\nu_2 = n - n_I$. The critical value for which the hypothesis of a null

factor effect is rejected at a significance level of p_c is then $F_c = F_{\nu_1,\nu_2,p_c}$, where $\nu_1 = n_I - 1$. Note that the inequality $\mathbb{E}[F] \geq F_c$ is satisfied for $\Delta \geq \frac{\nu_2}{\nu_2 - 2}\frac{F_c}{n_J} - \frac{1}{n_J}$. For a modestly sized experiment of $n_I = 10, n_J = 25$, with significance level $p_c = 0.05$, this yields $\Delta \geq 0.0361$, indicating that the magnitude of variability from errors may be nearly up to 30 times larger ($\Delta^{-1} \leq 27.7$) than variability from the factor of interest, but that it is not unlikely that the factor of interest will be deemed statistically significant. It will be rejected with a probability of 0.4434 when $\sigma_\epsilon^2 = \Delta^{-1}$. Table 1.3 displays values of Δ^{-1}, such that for variability from errors up to $\sigma_\epsilon^2 \leq \Delta^{-1}$, with $\sigma_\alpha^2 = 1$, the result will be $\mathbb{E}[F] \geq F_c$. The table further provides probabilities, $\Pr(F \geq F_c)$, that factor α will be deemed as statistically significant when using a significance level of 0.05. Geinitz and Furrer (2013) provide a more in-depth example in which two distinct scenarios of practical significance, which yield identical p-values, are compared.

Table 1.3: For given n_I, n_J, with $\sigma_\alpha^2 = 1$, if $\sigma_\epsilon^2 \leq \Delta^{-1}$, then $\mathbb{E}[F] \geq F_c$, where F_c is the critical value corresponding to a 0.05 level of significance for the hypothesis of a null factor effect. Entries shown in each cell of the table are then $\Delta^{-1}_{(p_{\text{rej}})}$, where $p_{\text{rej}} = \Pr(F \geq F_c)$ is the probability that the hypothesis is rejected for $\sigma_\epsilon^2 = \Delta^{-1}$.

		n_J				
		3	5	10	25	100
n_I	3	$1.2_{(0.3156)}$	$2.2_{(0.3668)}$	$4.7_{(0.3960)}$	$12.3_{(0.4075)}$	$49.9_{(0.4123)}$
	10	$2.6_{(0.4053)}$	$4.9_{(0.4286)}$	$10.6_{(0.4390)}$	$27.7_{(0.4434)}$	$112.9_{(0.4466)}$
	100	$9.7_{(0.4712)}$	$18.0_{(0.4749)}$	$38.5_{(0.4800)}$	$99.8_{(0.4811)}$	$406.3_{(0.4807)}$

The multilevel modeling community has also recognized that standard errors of the factor level predictors are also underestimated using classical methods, which then lead to conclusions of statistical significance (Goldstein, 1995). On a wider scale, across studies, the issue can be further compounded. Ioannidis (2005) discusses how many research findings are falsely identified based on statistical power, not only in individual studies, but also as well as a consequence of the number of investigators conducting similar research.

From a less tangible point of view, the ideological issue is primarily that the fundamental statistical method of ANOVA culminates with a test of significance, which is not always the most informative result. Researchers in other fields have consistently recommended confidence intervals and estimation over hypothesis testing (Yoccoz, 1991; Fidler *et al.*, 2004), although the results of standard ANOVA remain a default step in many analyses.

Mixed Model Controversy

Knowing when to regard a factor as fixed or as random is not always clear, as the definitions of fixed and random alone are not agreed upon by statisticians. Searle *et al.* (1992), Kreft *et al.* (1998), and Snijders and Bosker (1999) each offer different definitions. Gelman (2005) reviews these and others. It is also possible that interest lies in both the estimates of the individual α_i and in σ_α^2. Searle *et al.* (1992) pragmatically recommend the random effects model to be used whenever the additional assumption is deemed reasonable, and fixed otherwise.

Further problems arise when more terms are present in the statistical model. For $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, with α fixed and β random and interaction term $(\alpha\beta)$, it is not entirely clear how the identifiability constraints should influence the distribution of the random components β_j. Herein lies the so-called mixed models controversy (Voss, 1999). This issue leads to disagreements of which sums of squares quantities should be used in the test of a significance for the main random effect. Discussions in the literature on this issue have prolonged. Nelder (1977) has reiterated his views on parameter constraints in a series of papers (Nelder, 1994; Nelder *et al.*, 1998; Nelder, 2008), and has led authors to reconsider their previous positions, as with (Voss, 1999, 2000) and (Lencina *et al.*, 2005; Lencina and Singer, 2006). Due to the continued debate, one may assume that a universally agreed upon solution has not yet been found.

1.2 Hierarchical Bayesian ANOVA

In this section an overview of the steps in our proposed solution to the previously discussed issues is given. Geinitz and Furrer (2013) elaborate further on this and provide illustrative examples of a Bayesian ANOVA approach that may serve as a general strategy for ANOVA methods.

To resolve the issues discussed in the previous section there are two steps. The first step is to treat all factors in the model as random, as recommended by Gelman (2005). Taking this yet further, we opt for an explicitly hierarchical Bayesian model, assigning priors to all variance parameters in the statistical model. The second step is to use implicitly constrained prior distributions to avoid the issues that arise with the mixed model controversy.

The hierarchical nature of the model facilitates estimation of the magnitude of the variability that a factor contributes to overall variation in the data and for explicit comparison across all factors in the model, including errors. For example, observations modeled as $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, and a Bayesian parameterization for mean μ, factor α, and errors ϵ lead to the model specification

$$\mu \mid \sigma_\mu^2 \sim \mathrm{N}(\mu_0, \sigma_\mu^2), \qquad \alpha_i \mid \sigma_\alpha^2 \overset{iid}{\sim} \mathrm{N}(\alpha_0, \sigma_\alpha^2), \qquad \epsilon_{ij} \mid \sigma_\epsilon^2 \overset{iid}{\sim} \mathrm{N}(\epsilon_0, \sigma_\epsilon^2),$$
$$\sigma_\mu^2 \sim F_\mu, \qquad \sigma_\alpha^2 \sim F_\alpha, \qquad \sigma_\epsilon^2 \sim F_\epsilon,$$

where F_θ denotes an appropriate variance parameter distribution, e.g. inverse-gamma. More generally, consider each term in the model as a factor consisting of one or more levels, or a batch of effects. For the general case the β will be used. The terminology can be misleading, as the term *level* is not related to multilevel models. Then, following the notation of Gelman (2005), observations $Y_i, i = 1, \ldots, n$, are stated in terms of an additive decomposition

$$Y_i = \sum_{b=0}^{B} \beta_{j_i^b}^{(b)}. \tag{1.5}$$

A specific factor is indexed by $b = 0, \ldots, B$, with each factor consisting of n_b distinct levels. Factor indices $b = 0$ and $b = B$ typically correspond to overall mean, μ, and to measurement errors, ϵ_i, respectively, so that $n_0 = 1$, $n_B = n$. The individual levels of a factor are denoted by, $\beta_1^{(b)}, \ldots, \beta_{n_b}^{(b)}$, with $j_1^b, \ldots, j_n^b$ replicating them appropriately so that exactly one factor level from each factor is associated with exactly one observation. This may also be written explicitly as a linear model using a regression formulation

$$Y_i = \sum_{b=0}^{B} \sum_{j=1}^{n_b} x_{i,j}^{(b)} \beta_{j^b}^{(b)}, \tag{1.6}$$

with $\beta_{j^b}^{(b)}$ denoting the levels of a factor and $x_{i,j}^{(b)}$ denoting its corresponding linear predictors.

Since the Bth factor in the model typically refers to errors, or replications, it will be convenient to refer to this factor by ϵ and to treat it distinctly. Thus, the prior model specification for the levels of each of the $b = 0, \ldots, B-1$ factors in the model is given by

$$\beta_j^{(b)} \mid \sigma_b^2 \sim \mathrm{N}(\beta_0^{(b)},\ \sigma_b^2), \qquad j = 1, \ldots, n_b. \tag{1.7}$$

An appropriate variance parameter distribution, such as an inverse-gamma distribution, is chosen for the σ_b^2 terms. However, to maintain conjugacy, and to facilitate computation, prior distributions are of the form

$$\sigma_b^2 + \frac{n_b}{n} \sigma_\epsilon^2 \sim \mathrm{InvG}(u_b, v_b). \tag{1.8}$$

This specific prior parameterization is not novel as it has been used by Box and Tiao (1992) in a Bayesian variance components analysis setting, as well as by Everson and Morris (2000) in a multivariate hierarchical model setting.

The model specification is completed with the final factor, ϵ, and is given by

$$\epsilon_i \mid \sigma_\epsilon^2 \sim \mathrm{N}(0,\ \sigma_\epsilon^2), \qquad i = 1, \ldots, n, \tag{1.9}$$

$$\sigma_\epsilon^2 \sim \mathrm{InvG}(u_\epsilon, v_\epsilon). \tag{1.10}$$

Geinitz and Furrer (2013) derive a type of normal–inverse-gamma–inverse-gamma distribution that accounts for (1.7),(1.8), and (1.10), which allows for flexible parameterization while maintaining conjugacy.

The second step in our newly proposed ANOVA method is to work with implicitly constrained prior distributions for the factor levels. An appropriate constraint, e.g. $\mathbf{C}_b\boldsymbol{\beta}^{(b)} = a_b$, where $\boldsymbol{\beta}^{(b)} = (\beta_1^{(b)}, \ldots, \beta_{n_b}^{(b)})^{\mathrm{T}}$, is used and results in an improper form of (1.7). This type of joint normal distribution under a linear constraint can be considered as a special case of an intrinsic Gaussian Markov random field (Rue and Held, 2005). The primary advantage being that the full joint posterior may be conveniently factored when a balanced design is used. Assume a model of the form $Y_i = \beta^{(0)} + \beta_{j_i^1}^{(1)} + \cdots + \beta_{j_i^{B-1}}^{(B-1)} + \beta_i^{(B)}$. When least squares estimators that are independent of one another are available, then by placing similar constraints on the prior distributions as on the estimators, a practical factorization of the full joint posterior is possible. This is of the form

$$p(\sigma_\epsilon^2,\ \{\sigma_b^2, \beta^{(b)}\}_{b=0}^{B-1} \mid Y) = p(\sigma_\epsilon^2 \mid Y) \prod_{b=0}^{B-1} p(\sigma_b^2,\ \beta^{(b)} \mid Y,\ \sigma_\epsilon^2), \tag{1.11}$$

and allows for direct sampling from the posterior distribution without the need for computationally intensive MCMC sampling. Further explanation on the use of the constrained priors is introduced for the more general multivariate case in Section 2.1, and is developed further in Geinitz *et al.* (2012b).

As an applied procedure, this approach can be seen as a tool somewhere between linear mixed effects models and full MCMC routines. The former are fast, simple, and offer estimates of parameters, but typically give uncertainty of the estimates based on second-order approximations of the likelihood. The latter leads to more exact statements on the uncertainty of the estimates, but are computationally more intensive.

Although the data motivating our work is spatial in nature, this ANOVA framework can be applied to more general multivariate contexts as well.

2 Multivariate Hierarchical Bayesian ANOVA

The following section provides a brief explanation of the multivariate form of our hierarchical Bayesian approach to analysis of variance. Although multivariate hierarchical modeling in a

Bayesian context has previously been examined, e.g. Everson and Morris (2000), the inclusion of constraints on the prior distributions in this context are a novel contribution.

For multivariate data consisting of d elements the observations indexed by $i = 1, \ldots, n$ are each of dimension $\mathbb{R}^d$, and the analog of (1.5) is

$$\mathbf{Y}_i = \sum_{b=0}^{B} \boldsymbol{\beta}_{j_i^b}^{(b)}, \tag{2.1}$$

while for (1.6) it is

$$\mathbf{Y}_i = \sum_{b=0}^{B} \sum_{j=1}^{n_b} \mathbf{X}_{i,j}^{(b)} \boldsymbol{\beta}_{j^b}^{(b)}, \tag{2.2}$$

with $\mathbf{X}_{i,j}^{(b)} \in \mathbb{R}^{d \times d_b}$ having full column rank, and $\boldsymbol{\beta}_{j^b}^{(b)} \in \mathbb{R}^{d_b \times 1}$. The likelihood corresponding to (2.1) is then

$$\mathbf{Y}_i \mid \{\boldsymbol{\beta}_{j_i^b}^{(b)}\}_{b=0}^{B-1}, \boldsymbol{\Sigma}_\epsilon \sim \mathrm{N}_d \left(\sum_{b=0}^{B-1} \boldsymbol{\beta}_{j_i^b}^{(b)}, \ \boldsymbol{\Sigma}_\epsilon \right), \qquad i = 1, \ldots, n. \tag{2.3}$$

The prior model specification of the factor levels is then

$$\boldsymbol{\beta}_j^{(b)} \mid \boldsymbol{\Sigma}_b \sim \mathrm{N}_d(\boldsymbol{\beta}_0^{(b)}, \ \boldsymbol{\Sigma}_b), \qquad j = 1, \ldots, n_b, \tag{2.4}$$

and the multivariate form of (1.8) is

$$\boldsymbol{\Sigma}_b + \frac{n_b}{n} \boldsymbol{\Sigma}_\epsilon \mid \boldsymbol{\Sigma}_\epsilon \sim \mathrm{W}^{-1}(\boldsymbol{\Psi}_b, \ \kappa_b), \tag{2.5}$$

for $b = 0, \ldots, B-1$ and where $\mathrm{W}^{-1}(\boldsymbol{\Psi}, \kappa)$ refers to an inverse-wishart distribution with scale matrix $\boldsymbol{\Psi}$ and degrees of freedom κ. Distributions of error terms and of the error covariance are

$$\boldsymbol{\epsilon}_i \mid \boldsymbol{\Sigma}_\epsilon \sim \mathrm{N}_d(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon), \qquad i = 1, \ldots, n, \tag{2.6}$$

$$\boldsymbol{\Sigma}_\epsilon \sim \mathrm{W}^{-1}(\boldsymbol{\Psi}_\epsilon, \kappa_\epsilon). \tag{2.7}$$

Jeffreys' prior for the inverse-Wishart distribution is obtained for $\boldsymbol{\Psi} = \mathbf{0}, \kappa = 0$, so that the density is $p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(d+1)}$. For a reasonable number of observations this is possible, although using an identity scale matrix and $\kappa = d + 1$ does yield the uniform distribution for off-diagonal correlation parameters (Tokuda *et al.*, 2011), and can be useful in practice.

The likelihood and prior distributions are then combined to obtain a posterior distribution of all covariance parameters, and of the individual factor levels.

2.1 Posterior Factorization

To obtain a computationally convenient form of the posterior, improper prior distributions are used on the joint prior distribution of levels within a factor similarly to the univariate case.

In the linear model, estimating levels of factor b, i.e. $\{\boldsymbol{\beta}_1^{(b)}, \ldots, \boldsymbol{\beta}_{n_b}^{(b)}\}$, requires constraints. As before, including analogous constraints in the prior of a Bayesian model allows for convenient posterior factorization. Degrees of freedom for each factor are then accounted for in the corresponding covariance posterior. Using a vectorized form of the model, $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_n^T)^T \in \mathbb{R}^{nd}$ and $\boldsymbol{\beta}^{(b)} = (\boldsymbol{\beta}_1^{(b)^T}, \ldots, \boldsymbol{\beta}_{n_b}^{(b)^T})^T \in \mathbb{R}^{n_b d}$, is convenient for the development. The constraint $\mathbf{C}_b\boldsymbol{\beta}^{(b)} = \mathbf{0} \in \mathbb{R}^{c_b d}$, where there are c_b constraints, combined with (2.3), (2.4), is now

$$\mathbf{Y} \mid \{\boldsymbol{\beta}^{(b)}, \mathbf{C}_b\boldsymbol{\beta}^{(b)} = \mathbf{0}\}_{b=0}^{B-1}, \boldsymbol{\Sigma}_\epsilon \sim \mathrm{N}_{nd}\left(\sum_{b=0}^{B-1} \boldsymbol{\beta}^{(b)},\ \mathbf{I}_n \otimes \boldsymbol{\Sigma}_\epsilon\right), \tag{2.8}$$

$$\boldsymbol{\beta}^{(b)} \mid \mathbf{C}_b\boldsymbol{\beta}^{(b)} = \mathbf{0}, \boldsymbol{\Sigma}_b \sim \mathrm{N}_{n_b d}(\mathbf{1}_{n_b} \otimes \boldsymbol{\beta}_0^{(b)},\ \widetilde{\boldsymbol{\Omega}}_b), \qquad b = 0, \ldots, B-1, \tag{2.9}$$

with $\otimes$ denoting the Kronecker product. The singular matrix $\widetilde{\boldsymbol{\Omega}}_b$ has rank $d(n_b - c_b)$. To derive the improper prior the unconstrained and vectorized form of (2.4) is first used. This has covariance $\boldsymbol{\Omega}_b = \mathbf{I}_{n_b} \otimes \boldsymbol{\Sigma}_b$.

The density can also be stated in terms of the precision matrix, which is the inverse of the covariance, $\mathbf{Q}_b = \boldsymbol{\Omega}_b^{-1}$, and is also given as a Kronecker product. Using an eigendecomposition $\mathbf{Q}_b = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T \otimes \boldsymbol{\Sigma}_b^{-1}$, it is seen that this allows for arbitrary cases, other than the identity $\mathbf{I}_{n_b}$, to be used. The rank deficiency is introduced by removing c_b eigenvalues from the diagonal matrix $\boldsymbol{\Lambda}$, e.g. $\widetilde{\boldsymbol{\Lambda}} = \mathrm{diag}(0, \ldots, 0, \lambda_{c_b+1}, \ldots, \lambda_{n_b})$, leading to $\widetilde{\mathbf{Q}}_b = \boldsymbol{\Gamma}\widetilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}^T \otimes \boldsymbol{\Sigma}_b^{-1}$ (Rue and Held, 2005). Let $|\mathbf{Q}|_*$ be the pseudo determinant of a singular matrix, that is, the product of its non-zero eigenvalues. In the case of the identity, as in (2.8), the eigenvalues are one. Thus, densities of the likelihood (2.8) and of factor level priors (2.9) are

$$\begin{aligned} p(\mathbf{Y} \mid \{\boldsymbol{\beta}^{(b)}\}_{b=0}^{B-1}, \boldsymbol{\Sigma}_\epsilon) \propto |\boldsymbol{\Sigma}_\epsilon|^{-n/2} \exp\Bigg(& -\frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \widehat{\mathbf{Y}}_i)^T\boldsymbol{\Sigma}_\epsilon^{-1}(\mathbf{Y}_i - \widehat{\mathbf{Y}}_i) \\ & -\frac{1}{2}\sum_{b=0}^{B-1}\sum_{j=1}^{n_b}(\boldsymbol{\beta}_j^{(b)} - \widehat{\boldsymbol{\beta}}_j^{(b)})^T\frac{n}{n_b}\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{\beta}_j^{(b)} - \widehat{\boldsymbol{\beta}}_j^{(b)})\Bigg), \\ p(\boldsymbol{\beta}^{(b)} \mid \boldsymbol{\Sigma}_b) \propto |\widetilde{\mathbf{Q}}_b|_*^{1/2} \exp & \left(-\frac{1}{2}(\boldsymbol{\beta}^{(b)} - \mathbf{1} \otimes \boldsymbol{\beta}_0^{(b)})^T\widetilde{\mathbf{Q}}_b(\boldsymbol{\beta}^{(b)} - \mathbf{1} \otimes \boldsymbol{\beta}_0^{(b)})\right) \\ \propto |\boldsymbol{\Sigma}_b|^{-(n_b - c_b)/2} \exp & \left(-\frac{1}{2}\sum_{j=c_b+1}^{n_b}(\boldsymbol{\beta}_j^{(b)} - \boldsymbol{\beta}_0^{(b)})^T\boldsymbol{\Sigma}_b^{-1}(\boldsymbol{\beta}_j^{(b)} - \boldsymbol{\beta}_0^{(b)})\right), \end{aligned}$$

where $\widehat{\cdot}$ denotes least squares estimates. As before, when least squares estimates are independent of one another, then analogous constraints placed on factor level priors allow the posterior to

be factored as

$$p(\boldsymbol{\Sigma}_\epsilon,\ \{\boldsymbol{\Sigma}_b, \boldsymbol{\beta}^{(b)}\}_{b=0}^{B-1} \mid \mathbf{Y}) = p(\boldsymbol{\Sigma}_\epsilon \mid \mathbf{Y}) \prod_{b=0}^{B-1} p(\boldsymbol{\Sigma}_b,\ \boldsymbol{\beta}^{(b)} \mid \mathbf{Y},\ \boldsymbol{\Sigma}_\epsilon). \tag{2.10}$$

Each joint density of a set of factor levels and of the variance component for the factor, i.e. $p(\boldsymbol{\Sigma}_b, \boldsymbol{\beta}^{(b)} \mid \mathbf{Y}, \boldsymbol{\Sigma}_\epsilon)$, is then further partitioned. Using the Woodbury identity, and similar matrix identities, the formula

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{s})^T\mathbf{U}^{-1}(\boldsymbol{x}-\boldsymbol{s}) + (\boldsymbol{x}-\boldsymbol{t})^T\mathbf{V}^{-1}(\boldsymbol{x}-\boldsymbol{t}) \\ = \ \mathrm{tr}\left((\mathbf{U}+\mathbf{V})^{-1}(\boldsymbol{s}-\boldsymbol{t})(\boldsymbol{s}-\boldsymbol{t})^T\right) \ + \ (\boldsymbol{x}-\mathbf{P}^{-1}\boldsymbol{m})^T\mathbf{P}(\boldsymbol{x}-\mathbf{P}^{-1}\boldsymbol{m}),\end{aligned} \tag{2.11}$$

is derived, where $\mathbf{P} = \mathbf{U}^{-1}+\mathbf{V}^{-1}$, $\boldsymbol{m} = \mathbf{U}^{-1}\boldsymbol{s}+\mathbf{V}^{-1}\boldsymbol{t}$. The superpopulation covariance posterior and the factor level posteriors are then found through the decomposition of quadratic forms of factor levels and of least squares estimates

$$\begin{aligned}\mathrm{tr}\left[(\boldsymbol{\Sigma}_\beta + \frac{n_b}{n}\boldsymbol{\Sigma}_\epsilon)^{-1}\mathbf{B}_b\right] + \sum_{j=1}^{c_b}(\boldsymbol{\beta}_j^{(b)} - \widehat{\boldsymbol{\beta}}_j^{(b)})^T \frac{n}{n_b}\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{\beta}_j^{(b)} - \widehat{\boldsymbol{\beta}}_j^{(b)}) \\ + \sum_{j=c_b+1}^{n_b}(\boldsymbol{\beta}_j^{(b)} - \mathbf{P}_b^{-1}\boldsymbol{m}_j^{(b)})^T\mathbf{P}_b(\boldsymbol{\beta}_j^{(b)} - \mathbf{P}_b^{-1}\boldsymbol{m}_j^{(b)}),\end{aligned}$$

with $\mathbf{P}_b = \boldsymbol{\Sigma}_b^{-1} + \frac{n}{n_b}\boldsymbol{\Sigma}_\epsilon^{-1}$, $\boldsymbol{m}_j^{(b)} = \frac{n}{n_b}\boldsymbol{\Sigma}_\epsilon^{-1}\widehat{\boldsymbol{\beta}}_j^{(b)} + \boldsymbol{\Sigma}_b^{-1}\boldsymbol{\beta}_0^{(b)}$, and where $\mathrm{tr}(\cdot)$ denotes the trace operator. Additionally, $\mathbf{B}_b = \sum_{j=c_b+1}^{n_b}(\widehat{\boldsymbol{\beta}}_j^{(b)} - \boldsymbol{\beta}_0^{(b)})(\widehat{\boldsymbol{\beta}}_j^{(b)} - \boldsymbol{\beta}_0^{(b)})^T$, is analogous to a matrix sums of squares of the unconstrained factor level estimates that has been adjusted by the prior mean. This allows for the posterior (2.10) to be further simplified as

$$p(\boldsymbol{\Sigma}_\epsilon,\ \{\boldsymbol{\Sigma}_b,\ \boldsymbol{\beta}^{(b)}\}_{b=0}^{B-1} \mid \mathbf{Y}) = p(\boldsymbol{\Sigma}_\epsilon \mid \mathbf{Y}) \prod_{b=0}^{B-1} p(\boldsymbol{\Sigma}_b \mid \mathbf{Y},\ \boldsymbol{\Sigma}_\epsilon)\ p(\boldsymbol{\beta}^{(b)} \mid \mathbf{Y},\ \boldsymbol{\Sigma}_\epsilon,\ \boldsymbol{\Sigma}_b), \tag{2.12}$$

such that there is no need for computationally intensive MCMC procedures. The corresponding distributions of (2.12) are

$$\boldsymbol{\Sigma}_\epsilon \mid \mathbf{Y} \ \sim W^{-1}\left(\boldsymbol{\Psi}_\epsilon + \sum_{i=1}^{n}(\mathbf{Y}_i - \widehat{\mathbf{Y}}_i)(\mathbf{Y}_i - \widehat{\mathbf{Y}}_i)^T,\ \kappa_\epsilon + n - \sum_{b=0}^{B-1} n_b\right), \tag{2.13}$$

$$\boldsymbol{\Sigma}_b + \frac{n_b}{n}\boldsymbol{\Sigma}_\epsilon \mid \mathbf{Y}, \boldsymbol{\Sigma}_\epsilon \ \sim W^{-1}\left(\boldsymbol{\Psi}_b + \mathbf{B}_b,\ \kappa_b + n_b - c_b\right), \tag{2.14}$$

$$\boldsymbol{\beta}_j^{(b)} \mid \mathbf{Y}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\Sigma}_b \ \sim \begin{cases} \mathrm{N}_d\left(\widehat{\boldsymbol{\beta}}_j^{(b)},\ \frac{n_b}{n}\boldsymbol{\Sigma}_\epsilon\right) & j = 1, \ldots, c_b, \\ \mathrm{N}_d\left(\mathbf{P}_b^{-1}\boldsymbol{m}_j^{(b)},\ \mathbf{P}_b^{-1}\right) & j = c_b+1, \ldots, n_b, \end{cases} \tag{2.15}$$

for $b = 0, \ldots, B-1$. To sample from the full posterior, a realization from (2.13) is drawn, then realizations from (2.14) for each factor, and finally realizations of the levels from (2.15) for each factor. Factor levels are adjusted according to constraints and sample variances are calculated for each factor to obtain a realization of each finite-population variance, thereby eliminating the need for MCMC techniques.

3 Quantifying Multivariate Variability

Some advantages of the hierarchical Bayesian ANOVA approach for univariate data are in being able to assess how each distinct source of variation in the model contributes to overall variability, as well as quantifying how these sources compare with one another. In the multivariate setting, however, such assessments and comparisons are not straightforward. To be able to quantify and compare the amount of variability that each factor contributes to overall variability, the variance components must be explicitly compared. In the univariate case this is a comparison of the magnitude between scalars, i.e. ascertaining whether $\sigma_b^2 > \sigma_b^{2\prime}$. In the multivariate case variance components are covariance matrices. The natural progression of comparisons would be $\mathbf{\Sigma}_b > \mathbf{\Sigma}_{b'}$, where $\mathbf{A} > \mathbf{B}$ implies that $\mathbf{A}-\mathbf{B}$ is positive definite and noting that proper covariance matrices are positive definite. This forms, however, only a partial order, thus this comparison is not always available. In such a case it is necessary to define adequate covariance criteria.

To facilitate comparison of covariance matrices from different sources of variability (factors), dimensionality reduction is used to introduce a total ordering. Specifically, a function that maps the space of covariance matrices to scalars is used, e.g. $f\colon \mathbb{W}^{d\times d} \to \mathbb{R}$, where the codomain is typically restricted to positive values. Thus, the idea is that a type of criterion, or a set of criteria, is needed to quantify the amount of variability that is present in a covariance matrix. The objective of any criteria is to adequately summarize the qualities of the covariance that are of interest. Several such criteria are now given.

- *Effective variance* as defined by Pena and Rodriguez (2003) is

$$V_e(\mathbf{W}) = |\mathbf{W}|^{1/d}, \tag{3.1}$$

  although they use this as a measure of multivariate variability such that the notation is $V_e(\mathbf{X}) = |\mathbf{\Sigma}_\mathbf{X}|^{1/d}$, for $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \mathbf{\Sigma}_\mathbf{X})$.

- *Effective dependence* (Pena and Rodriguez, 2003) is defined as

$$D_e(\mathbf{W}) = 1 - |\mathbf{R}|^{1/d}. \tag{3.2}$$

  where $\mathbf{R}$ is the correlation matrix corresponding to $\mathbf{W}$, i.e. $\mathbf{R} = \mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}$, and $\mathbf{D}$ is the diagonal matrix with entries $\mathbf{D}_{i,i} = \sqrt{\mathbf{W}_{i,i}}$.

- *Sphericity* (Anderson, 1984, p427), used to test the hypothesis $H_0 : \mathbf{\Sigma} = \sigma^2\mathbf{I}$, is given by

$$\psi(\mathbf{W}) = \frac{|\mathbf{W}|^{1/d}}{(1/d)\mathrm{tr}(\mathbf{W})}. \tag{3.3}$$

  If $\psi = 1$, then the shape of the data with the given covariance is a sphere and H_0 is accepted. If $\psi = 0$, the data exists in a lower dimension, i.e. covariance is rank deficient. If the matrix is a correlation matrix, the relation $\psi(\mathbf{W}) = 1 - D_e(\mathbf{W})$ exists.

- *Generalized variance* (Mardia *et al.*, 1979) is a measure of the volume of the ellipsoidal contours of the distribution function, i.e. the determinant

$$V_g(\mathbf{W}) = |\mathbf{W}|. \tag{3.4}$$

- *Total variance* (Geinitz *et al.*, 2012b) is the sum of all the covariance matrix entries

$$V_t(\mathbf{W}) = \mathbf{1}^\mathsf{T}\mathbf{W}\mathbf{1}, \tag{3.5}$$

  The name *total variance* has previously been used to represent the trace of a covariance, i.e. the sum of the marginal variances (Mardia *et al.*, 1979).

- *Total marginal variance* (Geinitz *et al.*, 2012b) is the total of the marginal variances, which is the trace

$$V_m(\mathbf{W}) = \mathrm{tr}(\mathbf{W}). \tag{3.6}$$

  This criterion disregards covariance terms that are present and thus for a covariance matrix with off-diagonal summing to a strictly positive value, the identity $V_t(\mathbf{W}) > V_m(\mathbf{W})$ will be satisfied.

Since the covariance criteria (3.1)–(3.6) reduce the dimensionality of the random matrices, it is possible to consider their densities directly, as seen in Figure 3.1 for several different types of dependency structures. Quantiles are shown as dotted lines, means as dashed lines, and the criteria evaluated at the scale matrix $\mathbf{\Psi}$ is shown as a thick grey line. This illustrates the behavior of the different criteria as well as the uncertainty of their distribution for different $\mathbf{\Psi}$.

An advantage of (3.1)–(3.3) is that they are independent of dimension. However, comparisons to values that can be considered as the truth, i.e. the criterion of the scale matrix, $V_e(\mathbf{\Psi})$, are not as apparent as when the linear criteria of (3.5) and (3.6) are used. This is particularly noticeable when a covariance is nearly diagonal, as seen by the fact that the 0.025 and 0.975 quantiles often do not cover the true value for these criteria (bottom row of Figure 3.1). Alternatively, criteria (3.4)–(3.6) do provide coverage over the value corresponding to the true $\mathbf{\Psi}$. Again, because of the linearity of criteria (3.5) and (3.6), they correspond precisely to the underlying scalar parameters regardless of the form of the scale matrix (two right-most columns of Figure 3.1). While the first three criteria may be useful in certain cases, and can be used for hypothesis tests on the form of the covariance, the latter three are more appropriate for deriving confidence interval types of summaries.

Criteria (3.4)–(3.6) can be concisely summarized by inverse-χ^2_ν distributions, χ^{-2}_ν. For a matrix

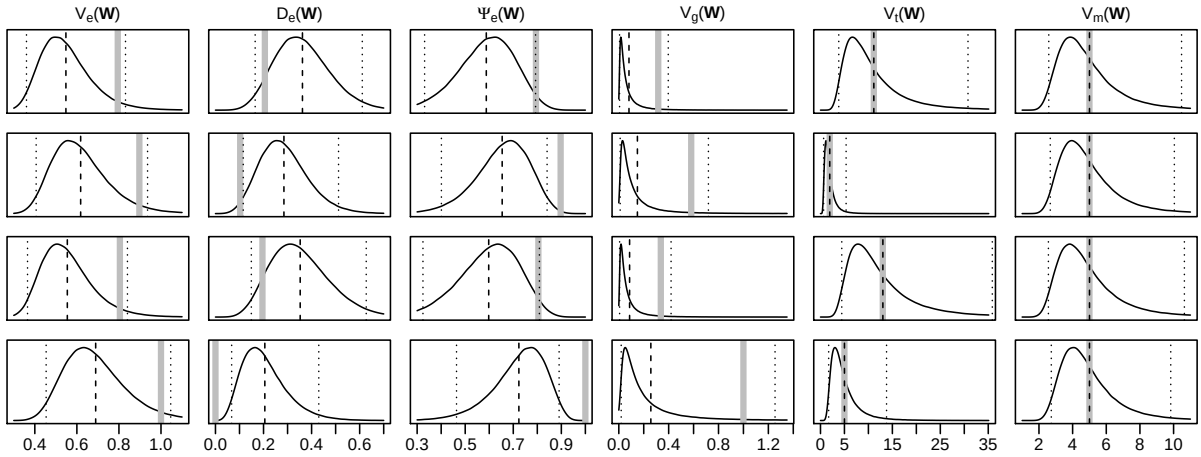

Figure 3.1: Comparison of scalar covariance criteria for a 5×5 matrix $\mathbf{W}$ that is inverse-Wishart distributed, $\mathrm{W}^{-1}(\mathbf{\Psi}, \kappa = 12)$, with scale matrix $\mathbf{\Psi}$ being equal to first-order autocorrelation matrix with correlation 0.5 (top row), first-order autocorrelation matrix with negative correlation -0.3 (second row), equicorrelation matrix with correlation 0.4 (third row), and identity matrix (bottom row). The mean (dashed vertical lines) and quantiles corresponding to 0.025 and 0.975 (dotted vertical lines) of the distribution are shown. The criteria functions evaluated at the scale matrix parameter, e.g. $V_e(\mathbf{\Psi})$, are denoted by thick grey lines.

parameter, $\mathbf{W}$, distributed as an inverse-Wishart, $\mathbf{W} \sim \mathrm{W}^{-1}(\Psi, \kappa)$, it is known that

$$V_g(\mathbf{W}) \sim |\mathbf{\Psi}| \cdot \left(\prod_{i=1}^{d} \chi^2_{\kappa-d+i} \right)^{-1}, \tag{3.7}$$

$$V_t(\mathbf{W}) \sim \mathbf{1}^T \mathbf{\Psi} \mathbf{1} \cdot \chi^{-2}_{\kappa-d+1}, \tag{3.8}$$

$$V_m(\mathbf{W}) \sim \sum_{i=1}^{d} \lambda_i \cdot \chi^{-2}_{\kappa-d+1}, \tag{3.9}$$

where $\lambda_1, \ldots, \lambda_d = \mathrm{eig}(\mathbf{\Psi})$ (Mardia *et al.*, 1979). In the example illustrated by Figure 3.1, functions are evaluated on $c\mathbf{W}$, with $c = \kappa - d - 1$. Since the expectation of an inverse-Wishart distribution is $\mathbf{\Psi}/c$, this facilitates direct comparison to the scale matrix.

Graphical representations are useful to understanding the results of any analysis, and in communicating results to others. However, visually displaying higher dimensional data and structures presents many challenges. Some examples of plotting higher dimensional data, are vector fields (e.g. wind data with speed and direction), ellipses or angles for correlation (Murdoch and Chow, 1996; Friendly, 2002; Wright, 2012), or glyphs. Visualizations of distributions of covariance matrices are even more challenging. Similar to the elliptical approach, these distributions can be presented by displaying a large set of ellipses that have been derived from the distribution in

which axis lengths correspond to eigenvalues and direction to eigenvectors of the decomposed covariance matrix. This is the approach used by Tokuda *et al.* (2011) and Geinitz *et al.* (2012b).

3.1 Spatial Data

Geostatistics is a subfield of statistics that is concerned primarily with data of a geographical, or spatial nature. The underlying assumption behind geostatistics is that geo-referenced data points that are close to one another are alike. As the distance between two points increases, then so does the degree of their independence (Cressie, 1993). Stationarity is also typically assumed, so that the covariance between two points s_1, s_2 does not depend on their locations, but only on the distance between them $h = \parallel s_1 - s_2 \parallel$. The advantage of such spatial dependence assumptions is that the covariance parameter space is reduced from $n(n+1)/2$ to only a few parameters, e.g. σ^2, ρ. Hence, spatial data is a type of multivariate data.

There exist many functions that define spatial dependence. Common among them are the parameters corresponding to a marginal variance and a range, denoted by σ^2 and ρ, respectively. The range governs how quickly spatial dependence dissipates and can be related to the concept of correlation, hence the use of the symbol ρ. The Matérn function is a general spatial covariance function and is defined as

$$C(h) = \sigma^2 \frac{1}{\Gamma(\nu)} \left(\frac{h}{2\rho}\right)^{\nu} 2K_{\nu}\left(\frac{h}{\rho}\right), \tag{3.10}$$

where ν is a smoothness parameter and K_{ν} is the modified Bessel function of the second kind (Matérn *et al.*, 1960). Setting $\nu = 0.5$ results in the exponential covariance, $C(h) = \sigma^2 \exp(-h/\rho)$. As $\nu \to \infty$ the Gaussian covariance, $C(h) = \sigma^2 \exp(-h^2/\rho)$, is obtained. While not a special case of the Matérn class of functions, the spherical covariance function,

$$C(h) = \begin{cases} \sigma^2, & \text{if } h = 0, \\ \sigma^2\left(1 - \frac{3h}{2\rho} + \frac{h^3}{2\rho^3}\right), & \text{if } 0 < h \leq \rho, \\ 0, & \text{if } h > \rho, \end{cases} \tag{3.11}$$

is also useful, as the covariance is zero at distances greater than range ρ (Schabenberger and Gotway, 2005). See also Wackernagel (2003) for a comprehensive review of spatial covariance functions that are available.

4 Higher Dimensions and Spatial Resolutions

For data in higher dimensions, such as the climate projections on the order of $d = 10^4$ in size that are examined herein, the steps needed for computation, i.e. obtaining posterior distribu-

tions, sampling from the posterior, etc., must be considered. Specifically, covariance matrices for multivariate data of this size would be approximately one-half gigabyte in size for 64-bit numerical representations.

Matrices of this size become prohibitively costly for operations needed in evaluating a density, or in generating a random sample. These types of computations require steps such as solving a linear system and calculating determinants. The standard approach to computation when large matrices are involved is to obtain a sparse matrix representation of covariances, so that the number of non-zero entries is on the order of d, rather than d^2. A Cholesky factorization is able to maintain the sparsity, thereby allowing for quick solving of a linear system and calculation of a determinant.

Several strategies exist to obtain sparse representations. For spatial processes with primarily local behavior, one can introduce sparsity through a tapering approach (Furrer *et al.*, 2006), i.e., direct multiplication of the covariance with a sparse correlation matrix. To adequately capture large scale structure Cressie and Johannesson (2008), and Stein (2008) have suggested exact methods for covariances essentially of the form $\mathbf{\Sigma} = \theta_0\mathbf{U} + \theta_1\mathbf{I}$, where $\mathbf{U}$ is an unknown matrix with rank p being small relative to its dimension. These lead to matrix operations dictated by the value of p. Rue and Tjelmeland (2002) take another approach, utilizing Gaussian Markov random field approximations to Gaussian processes which then result in sparse precision matrices. While the sparse precision matrix approach has been shown to produce reasonable approximations for spatial covariances (Lindgren *et al.*, 2011), it disregards the natural interpretation of spatial covariance functions and therefore this approach was not used here. For any of these strategies software libraries such as `spam` (Furrer and Sain, 2010) exist to carry out sparse matrix computations.

For the climate simulation data examined herein, estimates of spatial ranges have exceeded the limits for which a sparse matrix representation is possible. Practical values of sparsity (number of non-zero entries) for a matrix are dependent upon the operations being done. Table 4.1 shows the sparsity of the 8192×8192 covariance matrix, whose dimensions correspond to the 2.8° resolution shown in Figure 1, when various ranges over the Earth have been used. Timing of a matrix-vector multiplication is shown for both the sparse matrix object and the full matrix. This was carried out on a 2GHz Intel processor.

Due to the substantial ranges of the spatial fields considered, full covariance matrices are used. As a result, the use of approximations at other stages of the analysis have been implemented.

Table 4.1: Computation with a spatial covariance matrix over different ranges (column 1) in terms of memory storage (column 2) and processing time of a matrix-vector multiplication (column 3). Processing time for the multiplication with a full matrix representation is shown for comparison (column 4).

Range		Sparsity	Computation	Non-sparse Computation
(10^3 miles)	(degrees)	(% non-zero entries)	(10^{-3} seconds)	(10^{-3} seconds)
2	29°	8.3%	29	132
4	58°	25.8%	91	154
6	87°	47.4%	171	189
8	116°	69.6%	282	211
10	144°	88.4%	397	183

4.1 Approximating Marginal Variance Posteriors

Consider spatial data in which a factor α is distributed according a spatial covariance function and with errors ϵ being spatially independent, i.e. white noise. Observations reside in $\mathbb{R}^d$ and are given by

$$\mathbf{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij}, \quad i = 1, \ldots, n_\alpha, \quad j = 1, \ldots, n_\epsilon, \tag{4.1}$$

with model specification

$$\mathbf{Y}_{ij} \mid \boldsymbol{\mu}, \boldsymbol{\alpha}_i, \sigma_\epsilon^2 \sim \mathrm{N}_d(\boldsymbol{\mu} + \boldsymbol{\alpha}_i,\ \sigma_\epsilon^2 \mathbf{I}), \tag{4.2}$$

$$\boldsymbol{\alpha}_i \mid \sigma_\alpha^2 \sim \mathrm{N}_d(\boldsymbol{\alpha}_0,\ \sigma_\alpha^2 \mathbf{R}_\alpha), \tag{4.3}$$

$$u_\alpha = \sigma_\alpha^2 + \frac{n_\alpha}{n}\sigma_\epsilon^2 \mid \sigma_\epsilon^2 \sim \mathrm{InvG}(\psi_\alpha,\ \kappa_\alpha), \tag{4.4}$$

$$\sigma_\epsilon^2 \sim \mathrm{InvG}(\psi_\epsilon,\ \kappa_\epsilon), \tag{4.5}$$

for $i = 1, \ldots, n_\alpha$, $j = 1, \ldots, n_\epsilon$. As is sometimes done in analyses involving spatial fields, the range parameter is estimated a priori (Furrer *et al.*, 2007), as well as other smoothness parameters, e.g. ν for the Matérn function. Thus, the spatial correlation matrix $\mathbf{R}_\alpha$ is assumed to be known. The spectral decomposition $\mathbf{R}_\alpha = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T$, and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, yields the

superpopulation variance posterior densities

$$p(\sigma_\epsilon^2 \mid \mathbf{Y}) \propto p(\sigma_\epsilon^2) \cdot (\sigma_\epsilon^2)^{-(n-1-n_\alpha)/2} \exp\Big(-\frac{1}{\sigma_\epsilon^2}\frac{1}{2}\sum_{i=1}^{n_\alpha}\sum_{j=1}^{n_\epsilon}(\mathbf{Y}_{ij}-\widehat{\mathbf{Y}}_{ij})^T(\mathbf{Y}_{ij}-\widehat{\mathbf{Y}}_{ij})\Big), \quad (4.6)$$

$$\begin{aligned} p(\sigma_\alpha^2 \mid \mathbf{Y}, \sigma_\epsilon^2) &\propto p(\sigma_\alpha^2) \cdot |\sigma_\alpha^2 \mathbf{R}_\alpha + \frac{n_\alpha}{n}\sigma_\epsilon^2 \mathbf{I}|^{-(n_\alpha - c_\alpha)/2} \mathrm{etr}\Big(-\frac{1}{2}(\sigma_\alpha^2 \mathbf{R}_\alpha + \frac{n_\alpha}{n}\sigma_\epsilon^2 \mathbf{I})^{-1}\mathbf{S}_\alpha\Big) \\ &= p(\sigma_\alpha^2) \cdot \prod_{k=1}^{d} (\sigma_\alpha^2 \lambda_k + \frac{n_\alpha}{n}\sigma_\epsilon^2)^{\frac{-(n_\alpha - c_\alpha)}{2}} \exp\Big(-\frac{1}{2}\frac{(\mathbf{\Gamma}^T \mathbf{S}_\alpha \mathbf{\Gamma})_{k,k}}{\sigma_\alpha^2 \lambda_k + \frac{n_\alpha}{n}\sigma_\epsilon^2}\Big), \end{aligned} \quad (4.7)$$

where etr is the exponential trace operator and $\mathbf{S}_\alpha = \sum_{i=c_\alpha+1}^{n_\alpha} (\widehat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_0)(\widehat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_0)^\mathsf{T}$.

Using the form given in (4.7), the conditional posterior density $p(\sigma_\alpha^2 \mid \mathbf{Y}, \sigma_\epsilon^2)$ can be closely approximated by an inverse-gamma distribution. Note that each of the terms in the product are proportional to an inverse-gamma distribution that has been shifted. Using the substitution $u = \sigma_\alpha^2 \lambda_k + \frac{n_\alpha}{n}\sigma_\epsilon^2$, an inverse-gamma density $f_k(u) \propto u^{-s}\exp(-t_k/u)$, with $s = \frac{n_\alpha - c_\alpha}{2}$ and $t_k = \frac{1}{2}(\mathbf{\Gamma}^T \mathbf{S}_\alpha \mathbf{\Gamma})_{k,k}$, is found. The original support, $\sigma_\alpha^2 > -\frac{1}{\lambda_{\max}}\frac{n_\alpha}{n}\sigma_\epsilon^2$, where $\lambda_{\max}$ is the largest eigenvalue, is then $u > \frac{n_\alpha}{n}\sigma_\epsilon^2(1 - \frac{\lambda_k}{\lambda_{\max}})$.

As an illustration, consider an 8192×8192 spatial covariance matrix with range parameter $\rho = 4000$ miles. This resolution corresponds to geographical locations being 2.8° apart from one another. Climate data similar to that shown in Figure 1, and that is later examined (Furrer *et al.*, 2012), is utilized to obtain the required posterior parameters. The left and middle plots of Figure 4.1 illustrate a set of the individual shifted densities, f_k. The right side of the figure shows the original density (grey solid line) and the inverse-gamma approximation (black dotted line) that are found by using Laplace approximation to identify a normalizing constant, and first and second moments, and then moment matching to identify shape and scale parameters. It is often the case that approximations based on eigendecompositions of matrices fit reasonably well with a subset of f_k corresponding to the largest eigenvalues. However, because of the shifting and scaling due to both λ_k and t_k, a large number of functions f_k must be included. Hence, all are included in the approximation here, as there is no additional computational cost to do so.

4.2 Multivariate Gaussian Parameter Approximations

For additional factors in a spatial field analysis of variance procedure, a parameterization using (4.2)–(4.5) is utilized, where each factor in the model $b = 0, \ldots, B-1$ is considered to be in the form of (4.3), that is, a covariance matrix in the form of a spatial covariance. For prior mean $\boldsymbol{\alpha}_0 = \mathbf{0}$, the individual unconstrained factor levels for $i = c_\alpha + 1, \ldots, n_\alpha$ then have the posterior distribution

$$\boldsymbol{\alpha}_i \mid \mathbf{Y}, \sigma_\epsilon^2, \sigma_\alpha^2 \sim \mathrm{N}_d\left(\mathbf{P}^{-1}\boldsymbol{m}_i,\ \mathbf{P}^{-1}\right), \quad (4.8)$$

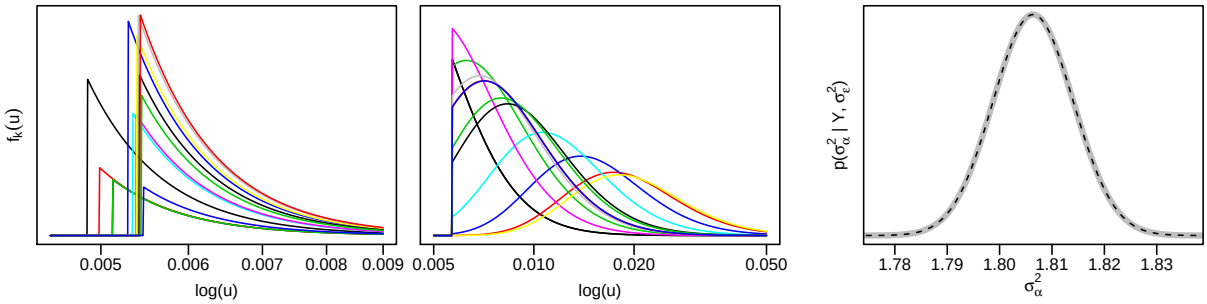

Figure 4.1: Individual densities in the product (4.7) (left and middle). These then yield an inverse-gamma approximation (right, grey solid line) for the posterior density of $\sigma_\alpha^2 \mid \mathbf{Y}, \sigma_\epsilon^2$ (right, black dotted line).

where $\mathbf{P} = (\sigma_\alpha^2 \mathbf{R}_\alpha)^{-1} + (\sigma_\epsilon^2 \frac{n_\alpha}{n} \mathbf{I})^{-1}$, $\boldsymbol{m}_i = \frac{n}{\sigma_\epsilon^2 n_\alpha} \widehat{\boldsymbol{\alpha}}_i$. Note that this is the form of (2.15), where $\boldsymbol{\Sigma}_b$ is a spatial covariance matrix. As a generalization consider a multivariate random variable

$$\mathbf{X} \sim \mathrm{N}_d \left(\boldsymbol{g},\ \mathbf{F} \right), \tag{4.9}$$

where $\mathbf{F} = \left((x_1 \mathbf{R})^{-1} + (x_2 \mathbf{I})^{-1} \right)^{-1}$ and $\boldsymbol{g} = \left((x_1 \mathbf{R})^{-1} + (x_2 \mathbf{I})^{-1} \right)^{-1} \frac{1}{x_2} \boldsymbol{s}$ are in the same form as $\mathbf{P}^{-1}$ and $\mathbf{P}^{-1} \boldsymbol{m}_i$, respectively. The mean and variance are thus considered as functions of two scalar variance components x_1, x_2, which are then considered as arguments to the d-dimensional vector valued function

$$\boldsymbol{g}(x_1, x_2) = \left(\frac{x_2}{x_1} \mathbf{R}^{-1} + \mathbf{I} \right)^{-1} \boldsymbol{s}, \tag{4.10}$$

and to the $d \times d$ matrix valued function

$$\mathbf{F}(x_1, x_2) = \left((x_1 \mathbf{R})^{-1} + (x_2 \mathbf{I})^{-1} \right)^{-1}. \tag{4.11}$$

For repeated sampling from (4.9) it is not efficient to require matrix inversion. Thus, a first-order series approximation of these functions is utilized, where the expansion of a vector/matrix valued function of two scalar parameters about the point (a_1, a_2) is

$$\mathbf{U}(x_1, x_2) \approx \mathbf{U}(a_1, a_2) + (x_1 - a_1) \frac{\partial \mathbf{U}}{\partial x_1}(a_1, a_2) + (x_2 - a_2) \frac{\partial \mathbf{U}}{\partial x_2}(a_1, a_2),$$

and yields the approximations

$$\begin{aligned}
\widetilde{\boldsymbol{g}}(x_1, x_2) &= \boldsymbol{g}(a_1, a_2) + (x_1 - a_1) \frac{1}{a_2} \mathbf{R} \left(\frac{a_1}{a_2} \mathbf{R} + \mathbf{I} \right)^{-2} \boldsymbol{s} - (x_2 - a_2) \frac{a_1}{a_2^2} \mathbf{R} \left(\frac{a_1}{a_2} \mathbf{R} + \mathbf{I} \right)^{-2} \boldsymbol{s} \\
&= \boldsymbol{g}_1 + (x_1 - a_1) \boldsymbol{g}_2 - (x_2 - a_2) \boldsymbol{g}_3, \\
\widetilde{\mathbf{F}}(x_1, x_2) &= \mathbf{F}(a_1, a_2) + (x_1 - a_1) a_2^2 \mathbf{R} \left(a_1 \mathbf{R} + a_2 \mathbf{I} \right)^{-2} + (x_2 - a_2) \frac{1}{a_2^2} \left(\frac{1}{a_1} \mathbf{R}^{-1} + \frac{1}{a_2} \mathbf{I} \right)^{-2} \\
&= \mathbf{F}_1 + (x_1 - a_1) \mathbf{F}_2 + (x_2 - a_2) \mathbf{F}_3,
\end{aligned}$$

where $\boldsymbol{g}_k, \mathbf{F}_k$, $k = 1, 2, 3$ may be computed one time so that scalar differences, $x_1 - a_1, x_2 - a_2$, are all that is needed for repeated sampling of $\widetilde{\mathbf{X}} \sim \mathrm{N}_d(\widetilde{\boldsymbol{g}}, \widetilde{\mathbf{F}})$. Of particular interest is the discrepancy incurred when sampling with $\widetilde{\mathbf{X}}$ as opposed to $\mathbf{X}$. That is, for standard multivariate normal random variable $\mathbf{Z}$, consider

$$\begin{aligned}\mathbf{X} &= \boldsymbol{g}(x_1, x_2) + \mathbf{F}(x_1, x_2)^{1/2}\mathbf{Z},\\ \widetilde{\mathbf{X}} &= \widetilde{\boldsymbol{g}}(x_1, x_2) + \widetilde{\mathbf{F}}(x_1, x_2)^{1/2}\mathbf{Z},\end{aligned}$$

where the square root denotes an appropriate matrix decomposition. The approximation discrepancy is measured using mean square error

$$\begin{aligned}\mathrm{MSE}[\widetilde{\mathbf{X}}] &= \mathbb{E}[(\mathbf{X} - \widetilde{\mathbf{X}})(\mathbf{X} - \widetilde{\mathbf{X}})^T]\\ &= (\boldsymbol{g} - \widetilde{\boldsymbol{g}})(\boldsymbol{g} - \widetilde{\boldsymbol{g}})^T + (\mathbf{F}^{1/2} - \widetilde{\mathbf{F}}^{1/2})(\mathbf{F}^{1/2} - \widetilde{\mathbf{F}}^{1/2})^T.\end{aligned}$$

Using the eigendecomposition $\mathbf{R} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T$, with $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda)$, this is further decomposed as

$$\begin{aligned}\mathrm{MSE}_g &= (\boldsymbol{g} - \widetilde{\boldsymbol{g}})(\boldsymbol{g} - \widetilde{\boldsymbol{g}})^T\\ &= (\boldsymbol{\Gamma}\mathbf{D}_g\boldsymbol{\Gamma}^T\boldsymbol{s})(\boldsymbol{\Gamma}\mathbf{D}_g\boldsymbol{\Gamma}^T\boldsymbol{s})^T, \qquad (4.12)\\ \mathrm{MSE}_F &= (\mathbf{F}^{1/2} - \widetilde{\mathbf{F}}^{1/2})(\mathbf{F}^{1/2} - \widetilde{\mathbf{F}}^{1/2})^T\\ &= \boldsymbol{\Gamma}\mathbf{D}_F\boldsymbol{\Gamma}^T. \qquad (4.13)\end{aligned}$$

Diagonal matrix $\mathbf{D}_g = \mathrm{diag}(d_g)$ is given by

$$\begin{aligned}\mathbf{D}_g &= \left(\frac{x_2}{x_1}\boldsymbol{\Lambda}^{-1} + \mathbf{I}\right)^{-1} - \left(\left(\frac{a_2}{a_1}\boldsymbol{\Lambda}^{-1} + \mathbf{I}\right)^{-1}\right.\\ &\qquad \left. + (x_1 - a_1)\frac{a_2}{a_1}\boldsymbol{\Lambda}^{-1}\left(\frac{a_2}{a_1}\boldsymbol{\Lambda}^{-1} + \mathbf{I}\right)^{-2} - (x_2 - a_2)\frac{1}{a_1}\boldsymbol{\Lambda}^{-1}\left(\frac{a_2}{a_1}\boldsymbol{\Lambda}^{-1} + \mathbf{I}\right)^{-2}\right)\\ &= \mathrm{diag}\left(\lambda\frac{(a_1x_2 - a_2x_1)}{(a_1\lambda + a_2)^2}\left(1 - \frac{a_1\lambda + a_2}{x_1\lambda + x_2}\right)\right), \qquad (4.14)\end{aligned}$$

while $\mathbf{D}_F = \mathrm{diag}(d_F)$ is

$$\begin{aligned}\mathbf{D}_F &= \mathbf{D} + \widetilde{\mathbf{D}} - 2\mathbf{D}^{1/2}\widetilde{\mathbf{D}}^{1/2}\\ &= \mathrm{diag}\left(\frac{x_1x_2\lambda}{x_1\lambda + x_2} + \lambda\frac{x_1a_2^2 + x_2a_1^2\lambda}{(a_1\lambda + a_2)^2} - 2\left(\frac{x_1x_2\lambda^2(x_1a_2^2 + x_2a_1^2\lambda)}{(a_1\lambda + a_2)^2(x_1\lambda + x_2)}\right)^{1/2}\right), \qquad (4.15)\end{aligned}$$

with

$$\begin{aligned}\mathbf{D} &= \left(\frac{1}{x_1}\boldsymbol{\Lambda}^{-1} + \frac{1}{x_2}\mathbf{I}\right)^{-1},\\ \widetilde{\mathbf{D}} &= \left(\frac{1}{a_1}\boldsymbol{\Lambda}^{-1} + \frac{1}{a_2}\mathbf{I}\right)^{-1} + (x_1 - a_1)a_2^2\boldsymbol{\Lambda}\left(a_1\boldsymbol{\Lambda} + a_2\mathbf{I}\right)^{-2} + \frac{x_2 - a_2}{a_2^2}\boldsymbol{\Lambda}^{-1}\left(\frac{1}{a_1}\boldsymbol{\Lambda}^{-1} + \frac{1}{a_2}\mathbf{I}\right)^{-2}.\end{aligned}$$

Recalling that x_1, x_2, while conditioned on, are random variables representing variance components in a hierarchical Bayesian ANOVA model, consider now the balanced case with $n = n_1 \cdot n_2$. That is, there are n_1 levels of factor 1 and n_2 replicates for each level such that factor 2 can be considered as the model error term. Random variables are then $x_1 = \sigma_1^2$, $x_2 = \sigma_2^2 \frac{n_1}{n} = \sigma_2^2 \frac{1}{n_2}$. Series expansion is done around the means of the respective parameters, $a_1 = \mu_1 = \mathbb{E}(x_1)$, and $a_2 = \frac{1}{n_2}\mu_2 = \frac{1}{n_2}\mathbb{E}(x_2)$ such that (4.14), (4.15) are

$$\mathbf{D}_g = \operatorname{diag}\left(\frac{\lambda}{n_2}\frac{(\mu_1\sigma_2^2 - \mu_2\sigma_1^2)}{(\mu_1\lambda + \mu_2\frac{1}{n_2})^2}\left(1 - \frac{\mu_1\lambda + \mu_2\frac{1}{n_2}}{\sigma_1^2\lambda + \sigma_2^2\frac{1}{n_2}}\right)\right) = \frac{1}{n_2}\mathbf{D}'_g,$$

$$\mathbf{D}_F = \operatorname{diag}\left(\frac{1}{n_2}\left(\frac{\lambda\sigma_1^2\sigma_2^2}{\sigma_1^2\lambda + \sigma_2^2} + \frac{\lambda(\sigma_1^2\frac{\mu_2^2}{n_2} + \sigma_2^2\mu_1^2\lambda)}{(\mu_1\lambda + \mu_2)^2} - 2\Big[\frac{\sigma_1^2\sigma_2^2\lambda^2(\sigma_1^2\frac{\mu_2^2}{n_2} + \sigma_2^2\mu_1^2\lambda)}{(\mu_1\lambda + \mu_2)^2(\sigma_1^2\lambda + \sigma_2^2)}\Big]^{1/2}\right)\right) = \frac{1}{n_2}\mathbf{D}'_F.$$

From which it follows that $\mathrm{MSE}[\widetilde{\mathbf{X}}]$ converges to 0 as $n_2 \to \infty$, since

$$\lim_{n_2\to\infty} \mathrm{MSE}_g = \lim_{n_2\to\infty} \frac{1}{n_2^2}\left(\mathbf{\Gamma}\mathbf{D}'_g\mathbf{\Gamma}^T\boldsymbol{s}\right)\left(\mathbf{\Gamma}\mathbf{D}'_g\mathbf{\Gamma}^T\boldsymbol{s}\right)^T = 0,$$

$$\lim_{n_2\to\infty} \mathrm{MSE}_F = \lim_{n_2\to\infty} \frac{1}{n_2}\mathbf{\Gamma}\mathbf{D}'_F\mathbf{\Gamma}^T = 0,$$

and implies convergence in probability, $\widetilde{\mathbf{X}} \xrightarrow{p} \mathbf{X}$, as well.

The MSE is now illustrated using data that has been analyzed in the thesis. Using an exponential spatial covariance matrix with range $\rho = 4000$ miles on a 2.8° resolution (same as that of Figure 1), surfaces of (4.12),(4.13), and of the total $\mathrm{MSE} = \mathrm{MSE}_g + \mathrm{MSE}_F$ are calculated. The values over which the surfaces are found are obtained using the joint densities of parameters $x_1 = \sigma_\alpha^2, x_2 = \frac{n_\alpha}{n}\sigma_\epsilon^2$, with distributions (4.7), (4.6), respectively. Note that the joint density is $p(x_1, x_2) = p(x_1|x_2)p(x_2)$, and that a negative correlation is typically present between the two. The data used to generate these posterior distributions is the same as that of Figure 4.1, and is later examined in more depth (Furrer *et al.*, 2012). The range over which the approximation is made is then $(x_1, x_2) \in [a_1 \pm 10 \cdot \mathrm{SD}(x_1)] \times [a_2 \pm 10 \cdot \mathrm{SD}(x_2)]$, where $\mathrm{SD}(\cdot)$ denotes the standard deviation and a_1, a_2 are the corresponding means. Results for MSE_g and MSE_F are shown in the left and middle plots of Figure 4.2. In the right-most plot, the total MSE is shown together with the (scaled) joint density of x_1, x_2, so that the area over which the approximation is being utilized can be seen. When this method is used in practice, (Furrer *et al.*, 2012; Geinitz *et al.*, 2012b), the support of the joint density has also been found to fall well within the bounds of area for which the linear approximation is accurate.

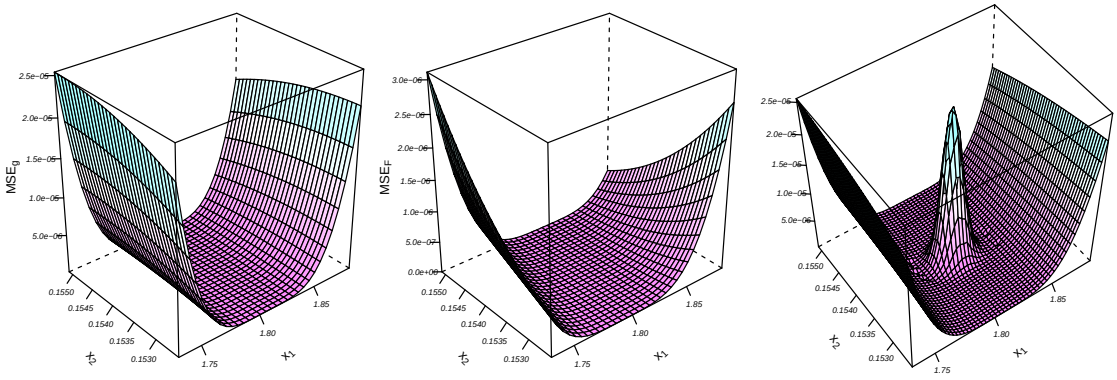

Figure 4.2: The mean-squared error, between a random variable $\mathbf{X}$ with mean $\boldsymbol{g}(x_1, x_2)$ and variance $\mathbf{F}(x_1, x_2)$ and $\widetilde{\mathbf{X}}$ with linearly approximated mean $\widetilde{\boldsymbol{g}}(x_1, x_2)$ and variance $\widetilde{\mathbf{F}}(x_1, x_2)$, is decomposed into the terms MSE_g (left) and MSE_F (middle). As x_1, x_2 are themselves random variables corresponding to marginal variances in a spatial covariance, the (scaled) joint distribution from an analysis of actual GCM data is shown together with total $\mathrm{MSE} = \mathrm{MSE}_g + \mathrm{MSE}_F$ (right), to see the area over which the approximation is utilized. Series approximation is done about the point (a_1, a_2), which is the mean of (x_1, x_2) and is located in the center of the plot.

Thesis Summary

The thesis consists of four manuscripts. The contents of these are now briefly summarized.

Paper I

Conjugate distributions in hierarchical Bayesian ANOVA for computational efficiency and assessments of both practical and statistical significance by Geinitz, S. and Furrer, R.

The motivation of this paper is to provide a summary of the current state of analysis of variance as a staple in the current corpus of statistical methodologies. There are shifts underway, towards methods that provide conclusions that are more relevant to scientific inquiries, but the procedures are not rigorously defined, and nor are computationally practical procedures widely available. In this paper a background of fixed, random, and mixed effects models is provided, and then put into the context of a hierarchical Bayesian model. Examples that highlight the ability of the new methodologies to identify practical significance of factors in a statistical model are also given. This manuscript is still partially under development but is intended to be submitted for publication in a suitable statistical journal.

The content and examples of the paper were completed by S. Geinitz.

Paper II

Multivariate analysis of global climate projections via rank deficient Bayesian ANOVA by Geinitz, S., Furrer, R., and Sain, S. R.

In this paper we introduce our hiearchical Bayesian approach to analysis of variance. The methodological contribution of the paper is in applying constrained prior distributions to factor levels so that there is no need for Gibbs, nor MCMC sampling. Although this is developed for the general multivariate case this can also be considered as a novel contribution for univariate data. Several scalar criteria of covariances as well as visualization tools for the distributions of covariance matrices are also discussed. The method is illustrated with a three-dimensional data in one-way ANOVA simulation and with bivariate data consisting of global climate averages for temperature and precipitation in a two-way ANOVA with temporal covariates. The climate data has been generated through climate simulations by roughly twenty atmospheric research

facilities and has been used in the Fourth IPCC Assessment Report. The two primary factors are the climate model and the emissions scenario employed. Despite the novel approach, our results are consistent with those presented in the IPCC report.

The contents and analyses of the paper were carried out by S. Geinitz. R. Furrer has, however, contributed through support and advising on all elements of the paper. S. Sain provided advice on the aspects related to the use of the climate data, and on the adequacy of the method for climate models.

Paper III

Bayesian analysis of variance for relative assessment of general circulation model effect
by Geinitz, S., Furrer, R., and Sain, S. R.

This paper further develops the hiearchical Bayesian analysis of variance method and applies it to large (128×64) spatial fields covering the globe. The method is extended to higher dimensional multivariate cases in which common spatial covariance functions are assumed, i.e. defined by range and marginal variance parameters. We approximate the posterior of the marginal variance by decomposing the correlation matrix through its eigendecomposition. This non-standard distribution can then be approximated by an inverse-gamma distribution to obtain the posterior distribution of the spatial fields. Simulations are done for varying dimensions of spatial fields. The method is then applied to simulated climate projections of temperature to assess uncertainty from climate model and from emissions scenario, both as constants and with respect to a linear temporal covariate. Graphical summaries of the posteriors show pointwise exceedance probabilities in order to illustrate how variability from the different factors compare over location.

The work in this paper has been carried out by R. Furrer and S. Geinitz. The simulation was conducted by S. Geinitz, as was much of the analysis of climate data. R. Furrer was responsible for much of climate discussion and selected figures of climate data, as well as for revisions of the paper. S. Sain provided input on initial modeling steps and also provided motivation through an earlier publication that was done in a similar spirit.

Paper IV

Assessing variance components of general circulation model output fields by Furrer, R., Geinitz, S., and Sain, S. R.

In this paper we have provided a more thorough analysis of the climate data that was part of the IPCC report by using spatial fields for both temperature and precipitation. Many graphical plots are provided in order to illustrate how climate projections vary over location due to the factors of climate model and emissions scenario, both as a constant and with a temporal covariate. Broadly speaking, the results show that climate model is responsible for the greatest amount of variation in the data, and that the emissions scenario does not have as strong of an affect. Posterior predictive distributions are shown to provide assessments of increases in temperature and precipitation over the globe in probabilistic statements. An early section of the paper provides an introduction and toy example of the hierarchical Bayesian approach to analysis of variance.

The writing and analyses in this paper were completed by S. Geinitz. R. Furrer has supervised on every step of the paper and offered suggestions for the most informative graphical summaries to provide. S. Sain reviewed the paper several times and gave input on the overall clarity of the paper.

Bibliography

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis, 2nd Edition.* Wiley-Interscience.

Bjerknes, V. (1904). *Das Problem der Wettervorhersage: Betrachtet vom Standpunkte der Mechanik und der Physik*, volume 1.

Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis.* Wiley-Interscience.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets.

Journal of the Royal Statistical Society. Series B (Statistical Methodology), **70**, 209–226.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons Inc.

Daley, R. (1993). *Atmospheric data analysis*, volume 2. Cambridge university press.

Everson, P. J. and Morris, C. N. (2000). Inference for multivariate normal hierarchical models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **62**, pp. 399–412.

Fidler, F., Geoff, C., Mark, B., and Neil, T. (2004). Statistical reform in medicine, psychology and ecology. *Journal of Socio-Economics*, **33**, 615–630.

Fisher, R. (1925). *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1 edition.

Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, **56**, 316–324.

Furrer, R., Geinitz, S., and Sain, S. R. (2012). Assessing variance components of general circulation model output fields. *Environmetrics*, **23**, 440–450.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, **15**, 502–523.

Furrer, R. and Sain, S. R. (2010). spam: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. *Journal of Statistical Software*, **36**, 1–25.

Furrer, R., Sain, S. R., Nychka, D. W., and Meehl, G. A. (2007). Multivariate Bayesian analysis of atmosphere-ocean general circulation models. *Environmental and Ecological Statistics*, **14**, 249–266.

Geinitz, S. and Furrer, R. (2013). Conjugate distributions in hierarchical bayesian anova for computational efficiency and assessments of both practical and statistical significance. Technical report, University of Zurich, Switzerland. URL http://arxiv.org/abs/1303.3390.

Geinitz, S., Furrer, R., and Sain, S. R. (2012a). Bayesian analysis of variance for relative assessment of general circulation model effect. *International Journal of Climatology (submitted)*.

Geinitz, S., Furrer, R., and Sain, S. R. (2012b). Multivariate analysis of global climate projections via rank deficient Bayesian ANOVA. *Journal of the Royal Statistical Society: Series C (under revision for resubmission)*.

Gelman, A. (2005). Analysis of variance: Why it is more important than ever. *Annals of Statistics*, **33**, 1–31.

Goldstein, H. (1995). *Multilevel Statistical Models*. Kendall's Library of Statistics. E. Arnold.

Houghton, J., Jenkins, G., Miller, H., and eds. (1990). Climate change, the ipcc scientific assessment.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Public Library of Science Medicine*, **2**, e124.

Kalnay, E. (2002). *Atmospheric modeling, data assimilation and predictability*, volume 1. Cambridge university press.

Kaufman, C. G. and Sain, S. R. (2010). Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis*, **5**, 847–874.

Kiehl, J., Hack, J., Bonan, G., Boville, B., Williamson, D., and Rasch, P. (1998). The national center for atmospheric research community climate model: CCM3*. *Journal of Climate*, **11**, 1131–1149.

Kreft, I., De Leeuw, J., and de Leeuw, J. (1998). *Introducing multilevel modeling*. Sage London.

Lencina, V. and Singer, J. (2006). Measure for measure: exact f tests and the mixed models controversy. *International statistical review*, **74**, 391–402.

Lencina, V. B., Singer, J. M., and Stanek, E. J. (2005). Much ado about nothing: the mixed models controversy revisited. *International Statistical Review*, **73**, 9–20.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 423–498.

Lorenz, E. (1972). Predictability: Does the flap of a butterflys wings in brazil set off a tornado in texas. American Association for the Advancement of Science.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press.

Matérn, B.*et al.* (1960). Spatial variation. stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden fran statens Skogsforskningsinstitut*, **49**, 1–144.

Murdoch, D. J. and Chow, E. D. (1996). A graphical display of large correlation matrices. *The American Statistician*, **50**, 178–180.

Nakićenović, N. and Swart, R., editors (2000). *Special Report on Emission Scenarios*. Intergovernmental Panel on Climate Change, Cambridge University Press.

Nelder, J.*et al.* (1998). The great mixed-model muddle is alive and flourishing, alas! *Food quality and preference.*, **9**, 157.

Nelder, J. A. (1977). A reformulation of linear models. *J. R. Statist. Soc. A*, **140**, 48–77.

Nelder, J. A. (1994). The statistics of linear models: back to basics. *Statistics and Computing*, **4**, 221–234.

Nelder, J. A. (2008). What is the mixed-models controversy? *International Statistical Review*, **76**, 134–135.

Nychka, D. W. (2000). Challenges in understanding the atmosphere. *Journal of the American Statistical Association*, **95**, 972–975.

Peixoto, J. and Oort, A. (1992). *Physics of climate*, volume 1. American Institute of Physics.

Pena, D. and Rodriguez, J. (2003). Descriptive measures of multivariate scatter and linear dependence. *Journal of Multivariate Analysis*, **85**, 361 – 374.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall, London.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, **29**, 31–49.

Schabenberger, O. and Gotway, C. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman and Hall.

Searle, S., Casella, G., and McCulloch, C. (1992). *Variance components*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Snijders, T. and Bosker, R. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Sage London.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K., Tignor, M., and Miller, H. (2007). Contribution of working group i to the fourth assessment report of the intergovernmental panel on climate change. *IPCC 2007 Climate Change: The Physical Science Basis*. Cambridge: Cambridge University Press.

Stein, M. L. (2008). A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society*, **37**, 3–10.

Tokuda, T., Goodrich, B., Van Mechelen, I., Gelman, A., and Tuerlinckx, F. (2011). Visualizing distributions of covariance matrices. Technical report, University of Leuwen, Belgium and Columbia University, USA. URL http://www.stat.columbia.edu/˜gelman/research/unpublished/Visualization.pdf.

Voss, D. T. (1999). Resolving the mixed models controversy. *American Statistician*, **53**, 352–356.

Voss, D. T. (2000). Resolving the mixed models controversy: Reply. *American Statistician*, **54**, 229–230.

Wackernagel, H. (2003). *Multivariate geostatistics*. Springer.

Wright, K. (2012). *corrgram: Plot a correlogram*. R package version 1.2.

Yoccoz, N. (1991). Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America*, **72**, 106–111.

Paper I

Conjugate distributions in hierarchical Bayesian ANOVA for computational efficiency and assessments of both practical and statistical significance

Steven Geinitz & Reinhard Furrer

Technical Report, University of Zurich, January 2013.

Conjugate distributions in hierarchical Bayesian ANOVA for computational efficiency and assessments of both practical and statistical significance

Steven Geinitz
geinitz@math.uzh.ch
University of Zurich
Zurich, Switzerland

Reinhard Furrer
furrer@math.uzh.ch
University of Zurich
Zurich, Switzerland

Abstract Assessing variability according to distinct factors in data is a fundamental technique of statistics. The method commonly regarded to as analysis of variance (ANOVA) is, however, typically confined to the case where all levels of a factor are present in the data (i.e. the population of factor levels has been exhausted). Random and mixed effects models are used for more elaborate cases, but require distinct nomenclature, concepts and theory, as well as distinct inferential procedures. Following a hierarchical Bayesian approach, a comprehensive ANOVA framework is shown, which unifies the above statistical models, emphasizes practical rather than statistical significance, addresses issues of parameter identifiability for random effects, and provides straightforward computational procedures for inferential steps. Although this is done in a rigorous manner the contents herein can be seen as ideological in supporting a shift in the approach taken towards analysis of variance.

Keywords: ANOVA; fixed effects; random effects; variance components; hierarchical Bayes; multilevel model; constraints

1 Introduction

As an independent field of study Statistics is rather young. Many of the methods, techniques, and philosophies can be attributed to a handful of statisticians during the first

half of the twentieth century. Among these, R.A. Fisher is often recognized as having had a profound influence on the field. It has been said that Fisher single-handedly created the foundations of modern statistical science (Hald, 1998). For statisticians the first contribution that comes to mind is his work in development of likelihood theory. However, for the greater scientific community, one might consider his formulation of analysis of variance as the most significant contribution.

As much of Fisher's work was in agriculture, an apt example to consider is the one-way ANOVA, $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, in which observations are on crop yield, with i, j representing the jth plant receiving fertilizer treatment i. An appropriate decomposition of the data should then reveal the variability due to different fertilizers while accounting for variability within plant groups that receive the same type of fertilizer treatment. Thus, analysis of variance is essentially a pragmatic decomposition of the data. In correspondence Fisher has been cited (Searle *et al.*, 1992) to have said,

> "The analysis of variance is (not a mathematical theorem but) a simple method of arranging arithmetical facts so as to isolate and display the essential features of a body of data with the utmost simplicity."

The elegance and power of this methodology is perhaps what has caused ANOVA to become so popular in nearly all areas of scientific research. However, along with the ubiquitous support of the methodology has come a pervasive reliance on its conclusory result, the p-value. Recognition of this problem is not new. It has been long noted by researchers in other fields that the hypothesis-based point of view, which relies on statistical significance, should be amended (Yoccoz, 1991; Fidler *et al.*, 2004; Ioannidis, 2005). The statistical community has also long acknowledged the need to provide methodologies that are first and foremost, "of use to scientists in making quantitative inferences," (Nelder, 1999). The problem is that the standard methods that continue to be imparted on students focus on statistical significance. As stated by Savage (1957), a method that does so "simply reflects the size of the sample and the power of the test, and is not a contribution to science." Thus, any standard, or default methodology that aims to decompose variation present in a set of observations according to factors of interest, should

be able to address practical significance as well.

In addition to the base objective of analysis of variance, to decompose variation in observations according to distinct sources of variability, a default method used in initial/exploratory work should accomplish the following.

- Allow for each factor to simultaneously consider variability due to the observed set of effects (finite population variance), as well as the variability from unobserved effects (superpopulation variance), thereby permitting greater flexibility in model choice with regards to fixed or random effects.

- Facilitate comparison of magnitude of variability across all factors in the model, including errors, so that attention may be given to practical significance as well as statistical significance of a factor.

- Provide ability to consider both magnitude and uncertainty of variance parameters in the model, by providing confidence, or uncertainty intervals in a default analysis summary.

These are precisely the goals of the analysis of variance framework proposed in this paper. While the primary contribution may be seen as ideological in nature, technical issues are also addressed, which help to facilitate the use of the more general framework.

The organization of the paper is as follows. Section 2 covers basic concepts of standard methods that are both widely taught and employed, as well as recent shifts in the practice of ANOVA. Section 3 presents an alternative framework of ANOVA along with modifications to the standard ANOVA table summary. Section 4 illustrates our method and compares it to the classical approaches. In particular, we present an example in which classical ANOVA yields identical p-values for two cases; one in which the factor under investigation has low practical significance, and one with high practical significance.

2 Background

Following Fisher's analysis of variance overall uncertainty is attributed to distinct factors of an experiment through the use of a sum of squares decomposition. This is now shown

with the balanced one-way analysis of variance model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad i = 1, \ldots, n_I, \quad j = 1, \ldots, n_J. \tag{1}$$

As a seminal example consider observations that are on crop yield with i, j representing the j^{th} plant receiving fertilizer treatment i. More generally the indices represent a factor level i and replicate j. The appropriate decomposition of the data, which reveals variability due to different fertilizers while accounting for variability within plant groups that receive the same type of fertilizer treatment, is done with the arithmetical arrangement that summarizes yield for each type, $\overline{Y}_{i\cdot} = n_J^{-1} \sum_j Y_{ij}$, and for overall yield, $\overline{Y}_{\cdot\cdot} = n^{-1} \sum_i \sum_j Y_{ij} = n_I^{-1} \sum_i \overline{Y}_{i\cdot}$, where $n = n_I \cdot n_J$. Observations Y_{ij} are decomposed with the identity

$$Y_{ij} - \overline{Y}_{\cdot\cdot} = (Y_{ij} - \overline{Y}_{i\cdot}) + (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}). \tag{2}$$

Terms are then squared and summed, noting that the cross term on the right hand side equals zero, so that a decomposition of the mean-adjusted sums of squares is

$$\underbrace{\sum_{i,j} (Y_{ij} - \overline{Y}_{\cdot\cdot})^2}_{\text{SST}} = \underbrace{\sum_{i,j} (Y_{ij} - \overline{Y}_{i\cdot})^2}_{\text{SSE}} + n_J \underbrace{\sum_{i} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2}_{\text{SSA}}, \tag{3}$$

The terms SST, SSA, and SSE denote *total (adjusted)* sum of squares, sum of squares *among* groups, and sum of squared *errors*, respectively. Note that each of these terms is itself a sum of squares that is analogous to a sample variance $s^2 = k^{-1} \sum_{i=1}^{k} (x_i - \overline{x})^2$, for a set of independent observations $x_1, \ldots, x_k$, and is thus proportional to a χ^2 distribution with appropriate degrees of freedom. Fisher showed that SSA and SSE are both proportional to χ^2 distributions, with $n_I - 1$ and $n - n_I$ degrees of freedom, respectively, and that they are independent, the general result of which is due to Cochran (1934).

While this classical methodology provides a means to examine statistical significance, it does not provide any formal assessment of practical significance. Loosely speaking, practical significance can be considered as a contextual basis that allows data-specific conclusions to be drawn, i.e. evidence that SSA is substantial when compared not only to zero, but to SSE as well. Practical significance in the example above implies that

the variability due to the fertilizer treatment is not only significantly different than no treatment, but that when compared to plant-to-plant variability it is still significant. One contribution of this paper is in attempting to formalize a statistical methodology that rigorously provides a method of assessing practical significance.

2.1 Conventional Methods

A fixed effects model generally refers to the case when the observations have exhausted the population of factor levels (e.g. treatments), or when interest lies only with the factor levels that have been observed. Alternatively, random effects models are employed when it is assumed that the factor levels are a subset of a greater population of possible levels. This definition provided by Hoaglin *et al.* (1991, p195) is somewhat more explicit than that given by Eisenhart (1947), in which the effects of a model are considered to be fixed when they are all nonrandom, and considered to be random when they are all random. There exist many other definitions in the literature, some of which are not compatible. See Gelman (2005) for a summary.

2.1.1 Fixed Effects

Consider the model given by (1) such that $i = 1, \ldots, n_I$ denotes the factor level or treatment, and $j = 1, \ldots, n_J$ denotes replications or errors. Observations are assumed to be independent across replicates as well as across factor levels. Additionally, it is generally assumed that

$$\epsilon_{ij} \sim N(0,\ \sigma_\epsilon^2). \tag{4}$$

Analysis of variance generally aims to test the hypothesis that there is no difference among the treatments,

$$H_0 : \alpha_1 = \cdots = \alpha_{n_I} = 0, \tag{5}$$

against the alternative hypothesis that at least one treatment level differs. The test is a result of the sums of squares decomposition in (3), since $\frac{\text{SSE}}{\sigma_\epsilon^2} \sim \chi^2_{n-n_I}$ and (under the null hypothesis) $\frac{\text{SSA}}{\sigma_\epsilon^2} \sim \chi^2_{n_I-1}$, where $n = n_I \cdot n_J$. The expectation of these two terms

is $\frac{n_J}{n_I-1}\sum_i \alpha_i^2 + \sigma_\epsilon^2$ and σ_ϵ^2, respectively. The test of H_0 is then carried out using the F distributed ratio $\frac{\text{MSA}}{\text{MSE}}$, where MSA $= \frac{\text{SSA}}{n_I-1}$ and MSE $= \frac{\text{SSE}}{n-n_I}$. The term MSA is central $\chi^2_{n_I-1}$ distributed when (5) is true, and non-central with shift of $\frac{n_J}{n_I-1}\sum_i \alpha_i^2 + \sigma_\epsilon^2$ when false.

The results described are concisely displayed in a tabular format (Fisher, 1925), as seen in Table 1. The table culminates with (5) being tested based on the p-value of $p = \Pr(F_{n_I-1,n-n_I} > F)$, which does not give any indication of the practical significance. And despite recognition of the need to focus on effect sizes and confidence intervals (Gardner and Altman, 1986; Nakagawa and Cuthill, 2007) rather than testing, the table remains a staple among statistical methodologies.

Table 1: One-way analysis of variance.

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Factor A	$n_I - 1$	SSA	MSA	$F = \frac{\text{MSA}}{\text{MSE}}$	$\Pr(F_{n_I-1,n-n_I} > F)$
Errors	$n - n_I$	SSE	MSE		

2.1.2 Random Effects

In addition to the statistical model (1) and distributional assumption (4), there is an additional assumption on the factor levels,

$$\alpha_i \sim N(0,\ \sigma_\alpha^2), \quad i = 1, \dots n_I, \tag{6}$$

with α_i and ϵ_{ij} independent. Observations are then normally distributed with mean and variance

$$\mathrm{E}[Y_{ij}] = \mu, \qquad \mathrm{Cov}(Y_{ij}, Y_{i'j'}) = \begin{cases} \sigma_\alpha^2 + \sigma_\epsilon^2 & i = i', j = j', \\ \sigma_\alpha^2 & i = i', j \neq j', \\ 0 & i \neq i'. \end{cases}$$

This parameterization has the added benefit that the parameter space for the factor levels is reduced from n_I to 1, since only σ_α^2 is estimated. Although individual levels, α_i, may

be predicted if necessary. Averaging over replications at factor level i yields the mean $\overline{Y}_{i\cdot}$, which are independently distributed $N(\mu, \sigma^2_{\alpha\epsilon})$, where $\sigma^2_{\alpha\epsilon} = \sigma^2_\alpha + \frac{\sigma^2_\epsilon}{n_J}$. Thus, the likelihood is a function of the three parameters μ, σ^2_ϵ, and σ^2_α.

Analogous to (5), the initial inquiry of interest is generally concerned with whether greater population variance σ^2_α is significantly different from zero. This corresponds to the null hypothesis

$$H_0 : \sigma^2_\alpha = 0, \tag{7}$$

and is tested using the same F-statistic as for (5) (Searle *et al.*, 1992; Rao, 1997; Cox and Solomon, 2003). Aside from its unintuitive nature, in that despite being random vs. fixed the same test statistic is used, this hypothesis test does little to remark on the practical significance of the variation due to factor α. Namely, the hypothesis may be rejected even when variation due to the errors is substantially greater, as seen in the example of Section 4.2.

Further inferential procedures on the variance components themselves are typically based on method of moments estimators, or explicitly use the likelihood. In the latter case, variability of the variance components are estimated with the Hessian of the likelihood, as with the widely used `R` packages `nlme` (Pinheiro *et al.*, 2006) and `lme4` (Bates and DebRoy, 2004). Wald-type confidence intervals can be then used to obtain confidence regions for the parameters. Similarly, the asymptotic properties of the log likelihood can be utilized to obtain confidence intervals using the χ^2 distribution, as seen in Figure 2 of Section 4.1.

2.1.3 Issues and Concerns

The choice to use a fixed or random effects model is not always immediately clear. The terminology alone may be seen as ambiguous since the distinction between fixed effects, random effects, and mixed effects is somewhat malleable. The simple fixed effects model of Section 2.1.1 can be seen as having a random component in the errors, ϵ_{ij}. Similarly, the random effects model of Section 2.1.2 can be seen as having a fixed component, μ. In both cases implying a mixed effects model. In practice a mixed effects model is employed

when there are two or more factors, other than overall mean and errors, and they are not all fixed (random).

More difficult perhaps is determining when each of these methods should be used. If interest lies in the distribution of the random effects, i.e. the variance component σ_α^2, then a random effects model should be chosen. If interest lies in the realized/observed levels of the factor, then a fixed effects model is used. If both are of interest, then the random effects should be chosen and levels are then predicted, rather than estimated. Searle *et al.* (1992, p18) take a pragmatic approach to this by recommending that in any case in which it is reasonable to assume that the levels of the factor come from a probability distribution, i.e. that (6) may be assumed, then a random effects model should be chosen. The usage of a random effects model, however, typically precludes the estimation of the finite population variance.

An additional problem that arises in analysis of variance with several factors is the so-called mixed models controversy (Voss, 1999; Lencina *et al.*, 2005; Nelder, 2008). The problem essentially comes down to how a hypothesis test of a random effect is carried out when an interaction is also present in the model.

To resolve the issues above we support the notion of Gelman (2005), in that *all* factors in the model are treated as random. The procedural steps are then carried out equivalently. If interest is in the observed (unobserved) levels of a factor, then greater focus is given to the finite (super) population variance. However, because of parameter dependencies involved in the unconstrained factor levels, Gelman recommends using MCMC, in which redundant parameterization is used in order to reduce dependencies and to speed up posterior sampling. Alternatively, we recommend using constraints to define an improper joint prior distribution on the factor levels, thereby eliminating the need for complex MCMC procedures, as in Geinitz *et al.* (2012). For the variables of interest and the scenarios examined herein, the use of improper distributions avoids the marginalization paradox of Dawid *et al.* (1973).

2.2 Multilevel Models

Often times the results of an analysis should allow for simultaneous consideration of both group level and individual level variability, e.g. variability according to schools and to students within schools. Applications of such scenarios initially arose in the social sciences (Goldstein, 1995; Kreft *et al.*, 1998; Snijders and Bosker, 2011), but have also included the health sciences (Von Korff *et al.*, 1992; Greenland, 2000), and have provided the basis for much of the work in multilevel models.

A multilevel model can be seen as a linear model with coefficients, i.e. factor levels, that are themselves modeled (Gelman and Hill, 2006). More generally, this can be considered as a type of hierarchical Bayesian approach. However, while not explicit, the multilevel point of view is useful in considering a generalized approach towards analysis of variance. Because the simultaneous consideration of group and individual level variability entails the decomposition of variation according to each of these sources,"ANOVA is fundamentally about multilevel modeling" (Gelman, 2005). That is to say, analysis of variance from the viewpoint of multilevel models allows for both finite population and superpopulation variance components to be considered, which can be seen as a unification of fixed and random effects. This comprehensive approach to analysis of variance yields useful results and has been used in other fields such as ecology (Qian and Shen, 2007), genetics (Leinonen *et al.*, 2008), and climate (Sain *et al.*, 2011).

In practice there have been some hindrances in the adoption of this more general approach to ANOVA. Computational procedures to carry out such an analysis typically rely on either mixed effects models (e.g. `lme4` package in `R`) or on MCMC methods (e.g. `WinBUGS`). However, while mixed effects models can be used to obtain initial estimates of the parameters in a multilevel model, inferential steps, e.g. confidence intervals, for variance parameters are often done through likelihood approximation. For more explicit inferential procedures it is necessary to use MCMC methods (Gelman and Hill, 2006, p566). Although the added complexity and computation of MCMC, particularly when the use is as an exploratory analysis step, can be a deterrent to this approach. A method that is both precise in its inferential statements while being straightforward to implement is not widely known.

2.3 Bayesian Results

The hierarchical approach towards analysis of variance can be explained most readily in a Bayesian framework. In an effort to explain this approach in a classical inference framework Gelman (2005) recommends a simulation approach, which is reminiscent of posterior sampling. Because we prefer to adopt an explicit Bayesian approach, we now review some results on distributions for variance components that facilitate the procedure.

2.3.1 Prior Distributions

For the normally distributed random variable $Y \sim \mathrm{N}(\mu, \sigma^2)$, prior specification of the parameters can be done in many different ways. Initially, consider μ, σ^2 to be either known or unknown, each in turn. Following the invariance principle (Jeffreys, 1946), prior distributions in univariate cases are then

$$\begin{aligned} \mu \text{ known, } \sigma^2 \text{ unknown:} \quad & p(\mu) \propto \text{const}, \\ \mu \text{ unknown, } \sigma^2 \text{ known:} \quad & p(\sigma^2) \propto (\sigma^2)^{-1}. \end{aligned}$$

Box and Tiao (1992, p43) derive same priors using the concept of location and scale parameters. These identical priors are also found using the reference approach of Bernardo and Smith (2000, p314), due to asymptotic normality of the posterior distributions. Note that the density of $p(\sigma^2) \propto (\sigma^2)^{-1}$ corresponds to an inverse-gamma distribution, $\Gamma^{-1}(u, v)$, with $u = v = 0$, where $x \sim \Gamma^{-1}(u, v)$ implies the density $p(x) \propto x^{-u-1} \exp(-\frac{v}{x})$. Common values of hyperparameters have thus been limiting forms thereof, such as $u = v = \varepsilon$, with ε small (Lunn *et al.*, 2000). If prior independence between μ and σ^2 is assumed, then the two univariate priors are combined for

$$p(\mu, \sigma^2) = p(\mu)p(\sigma^2) \propto (\sigma^2)^{-1}. \tag{8}$$

Alternatively, Jeffreys' prior for multivariate parameters $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ without independence leads to

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-3/2}. \tag{9}$$

These correspond to $\sigma^2 \sim \Gamma^{-1}(u, v)$, with $u = v = 0$ for the prior given by (8), and $u = \frac{1}{2}, v = 0$ for the prior given by (9).

Box and Tiao (1992, p251) decompose the likelihood by group means, e.g. $\overline{Y}_{i\cdot}$ in (2), to place a prior directly on $\sigma^2_{\alpha\epsilon}$. The joint prior distribution for $\mu, \sigma^2_\epsilon, \sigma^2_{\alpha\epsilon}$ is then

$$p(\mu, \sigma^2_\epsilon, \sigma^2_{\alpha\epsilon}) \propto (\sigma^2_\epsilon \sigma^2_{\alpha\epsilon})^{-1}. \tag{10}$$

Additionally, Jeffreys' independence prior of the original variance parameters $(\sigma^2_\epsilon, \sigma^2_\alpha)$ also leads to (10) (Box and Tiao, 1992). The multivariate analog of this has been used as well by Everson and Morris (2000). Naturally, a prior may also be placed directly on the parameter σ^2_α, although the posterior may no longer be as simple to work with.

2.3.2 Conjugacy

For observations $Y_i \sim \mathrm{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$, a multivariate conjugate prior for parameter $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ is the normal-inverse-gamma distribution. This is denoted as $\mathrm{N}\Gamma^{-1}(\mu_0, \tau, u, v)$ with $\mu_0 \in \mathcal{R}, \tau, u, v > 0$. More specifically,

$$\mu \mid \sigma^2 \sim \mathrm{N}(\mu_0,\ \frac{\sigma^2}{\tau}), \tag{11}$$

$$\sigma^2 \sim \Gamma^{-1}(u,\ v), \tag{12}$$

with joint density,

$$\begin{aligned} p(\mu, \sigma^2) &= p(\mu|\sigma^2) \cdot p(\sigma^2) \\ &= (2\pi\frac{\sigma^2}{\tau})^{-1/2} \exp\left(-\frac{\tau}{2\sigma^2}(\mu - \mu_0)^2\right) \cdot \frac{v^u}{\Gamma(u)}(\sigma^2)^{-u-1} \exp\left(-\frac{v}{\sigma^2}\right). \end{aligned} \tag{13}$$

Priors corresponding to (8) and (9) can be found through limiting forms of this. These are given, informally, by $\mathrm{N}\Gamma^{-1}(0, 0, -\frac{1}{2}, 0)$ and by $\mathrm{N}\Gamma^{-1}(0, 0, 0, 0)$, respectively. Conjugate priors of this form have been used extensively, although often with precision, $\tau = (\sigma^2)^{-1}$, resulting in a normal-gamma distribution (Bernardo and Smith, 2000, p136). The utility of this general parameterization is in being able to conform to different prior specifications while maintaining conjugacy. The full model is specified by the likelihood and prior

$$Y \mid \mu, \sigma^2 \sim \mathrm{N}(\mu,\ \sigma^2), \tag{14}$$

$$(\mu, \sigma^2) \sim \mathrm{N}\Gamma^{-1}(\mu_0,\ \tau,\ u,\ v), \tag{15}$$

respectively, with the posterior distribution given by

$$(\mu, \sigma^2) \mid Y \sim \mathrm{N}\Gamma^{-1}\left(\frac{\tau\mu_0 + n\overline{y}}{\tau + n},\, \tau + n,\, u + \frac{n}{2},\, v + \frac{1}{2}\left[\sum_i (y_i - \overline{y})^2 + \frac{(\overline{y} - \mu_0)^2}{n^{-1} + \tau^{-1}}\right]\right). \tag{16}$$

3 Comprehensive ANOVA

Following the view of Gelman (2005) we see the hierarchical Bayesian approach towards ANOVA (Section 2.2) as a means to unify the two distinct fixed and random effects models. In addition to the hierarchical model structure a Bayesian model specification is intuitive and practical. By following this approach the challenges discussed in Section 2.1.3 are resolved.

Hierarchical Bayesian models are typically considered simply as mixed effects models within the statistical community. However, because mixed effects models do not typically provide assessments of uncertainty of the variance component estimates, nor is variability of the observed set of factor levels examined by default, we do not see this as truly providing a comprehensive approach towards ANOVA. As stated, multilevel modeling seems to be a more natural strategy. As a result much of the work with multilevel models, including analysis of variance according to various factors, has largely taken place in other domains, primarily the social sciences (Goldstein, 1995; Gelman and Hill, 2006; Snijders and Bosker, 2011). This can be seen as a failure of the statisticians, as Huber (2011) states, "the consequences of not being able to adequately summarize and disseminate common methodologies may be a divergence of statistics, that each field develops its own version of statistics." By presenting ANOVA in a more general hierarchical framework we are also, "unifying the philosophies, concepts, statistical methods, and computational tools" (Lindsay *et al.*, 2004).

The unification of fixed and random effect models is clearly seen in the graphical model of Figure 1. The successive layers of distributional assumptions is shown clearly here. The inner-most box represents the fixed effects model, while the middle box represents the random effects model. The hierarchical Bayesian ANOVA model, or simply *comprehensive ANOVA*, is then represented by the outer-most box. The diagram explicitly shows the

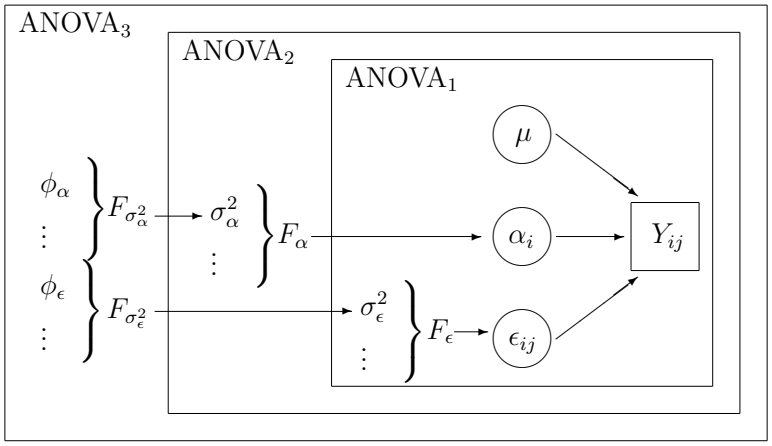

Figure 1: Graphical model representing successive assumptions for the fixed effect (inner box), random effect (middle box), and fully Bayesian (outer box) specifications.

unification of the models and immediately conveys a general view of ANOVA to students and researchers not familiar with variance analyses. The notation created by Eisenhart (1947) is here, where ANOVA_1 corresponds to M_1, the fixed effects model; and M_2 to ANOVA_2, the random effects model. Instead of M_3, which refers to a mixed effects model, we have chosen to allow ANOVA_3 to refer to a fully Bayesian parameterization. This can be confusing though, as Cox and Solomon (2003) have pointed out, "Occasionally the word Bayesian is used for any analysis involving more than one level of random variation." We agree with them, in that this can seem quite confusing, but nonetheless consider ANOVA_3 as a Bayesian analysis of variance procedure.

Analysis of variance in this framework allows the questions discussed in the Introduction to be addressed, and also resolves many of the issues discussed in Section 2.1.3. Gelman (2005) presents graphical summaries of this ANOVA approach that allow for visual comparison of confidence intervals for the variance components, which is possible for both finite and superpopulation variance parameters. In Table 2 a proposed alternative to the traditional ANOVA table is shown. Commonly significance in the classical ANOVA table is merely a function of power. That is, given enough observations, nearly any effect will be deemed as statistically significant. Alternatively, Table 2 provides estimates of the variance parameters, both finite and superpopulation, as well as a probabilistic

assessment of practical significance. This is done with a direct comparison of posterior distributions of all variance components against the error variance σ^2_ϵ. A probability regarding hypothesis (7), i.e. that the superpopulation variance σ^2_α is equal to zero, can be given as well. This probability, $\Pr(\sigma^2_\alpha = 0|Y)$, thus provides an assessment of statistical significance. While a point null hypothesis is typically not of interest in a Bayesian setting, there are two reasons that it is included here. First is to allow a comparison to the standard frequentist method of ANOVA, which relies on a point null hypothesis. Second is when a mass point is allowed in the posterior distribution, as will be seen in Section 4.

3.1 One-way Model

In the case of a single source of variation, as in (1), the model can be stated as $Y_{ij} = \alpha_i + \epsilon_{ij}$. To fully illustrate the novel approach towards ANOVA, what can be referred to as ANOVA_3, this basic experimental setup is all that is required.

3.1.1 Model Specification

A particularly useful parameterization in the one-way configuration, that allows different variance parameterizations while maintaining conjugacy, is an extension of the normal-inverse-gamma distribution. This involves an additional inverse-gamma distribution for

Table 2: Comprehensive ANOVA summary utilizing posterior distributions to obtain a summary of variance parameters (in units of standard deviation). Quantiles are used to provide a type of confidence interval. The probability $\Pr(\sigma_\alpha > \sigma_\epsilon|Y)$ provides an assessment on practical significance for the parameter.

	Parameter	Mean	Median	Uncertainty Interval	Sig. Rel. to Errors
α	(finite) s_α	$\mathrm{E}[s_\alpha\|Y]$	$Q_{0.5}[s_\alpha\|Y]$	$(Q_{0.025}[s_\alpha\|Y],\ Q_{0.975}[s_\alpha\|Y])$	$\Pr(s_\alpha > \sigma_\epsilon\|Y)$
	(super) σ_α	$\mathrm{E}[\sigma_\alpha\|Y]$	$Q_{0.5}[\sigma_\alpha\|Y]$	$(Q_{0.025}[\sigma_\alpha\|Y],\ Q_{0.975}[\sigma_\alpha\|Y])$	$\Pr(\sigma_\alpha > \sigma_\epsilon\|Y)$
ϵ	σ_ϵ	$\mathrm{E}[\sigma_\epsilon\|Y]$	$Q_{0.5}[\sigma_\epsilon\|Y]$	$(Q_{0.025}[\sigma_\epsilon\|Y],\ Q_{0.975}[\sigma_\epsilon\|Y])$	–

the added variance component. The resulting distribution is given by

$$\alpha_i \mid \alpha_0, \tau_\alpha, \tau_\epsilon, \sigma^2_\alpha, \sigma^2_\epsilon \sim \mathrm{N}\left(\alpha_0,\ \left[\frac{\tau_\alpha}{\sigma^2_\alpha} + \frac{\tau_\epsilon}{\sigma^2_\epsilon}\right]^{-1}\right), \tag{17}$$

$$\sigma^2_{\alpha\epsilon} \mid \sigma^2_\epsilon \sim \Gamma^{-1}(u_\alpha,\ v_\alpha), \tag{18}$$

$$\sigma^2_\epsilon \sim \Gamma^{-1}(u_\epsilon,\ v_\epsilon), \tag{19}$$

with $i = 1, \ldots, n_I$ corresponding to the number of groups, and additional parameter κ_ϵ such that $\sigma^2_{\alpha\epsilon} = \sigma^2_\alpha + \kappa_\epsilon \sigma^2_\epsilon$. Noting that $\sigma^2_{\alpha\epsilon}$ is analogous to the variance of mean of the observations sharing the same factor level, as in Section 2.1.2, it follows that κ_ϵ will typically be set to $\frac{1}{n_J}$. Hyperparameters $\tau_\alpha, \tau_\epsilon$ will often be $1, n_J$, respectively, although they are included here to allow for more flexible parameterizations and to comply with the form of the normal-gamma and normal-inverse-gamma distributions of Bernardo and Smith (2000).

The variance parameters and factor levels can then be jointly specified as a combination of normal and inverse gamma distributions, i.e. $\mathrm{N}\Gamma^{-1}\Gamma^{-1}(\alpha_0, \tau_\alpha, \tau_\epsilon, \kappa_\epsilon, u_\alpha, v_\alpha, u_\epsilon, v_\epsilon)$, with certain values of hyperparameters, or limits thereof, yielding different prior specifications such as those discussed earlier.

This is in general, however, not a conjugate model specification, i.e. the posterior distribution will not be of the same family as the prior distribution. Although the posterior σ^2_ϵ continues to follow an inverse-gamma distribution, with density

$$p(\sigma^2_\epsilon \mid Y) \propto (\sigma^2_\epsilon)^{-u_\epsilon - \frac{n-n_I}{2} - 1} \exp\left(-\frac{1}{\sigma^2_\epsilon}\left[v_\epsilon + \frac{1}{2}\sum_{ij}(y_{ij} - \widehat{\alpha}_i)^2\right]\right),$$

the posterior distribution of $\sigma^2_\alpha + \kappa_\epsilon \sigma^2_\epsilon$ will not be the same type as its prior for arbitrary values of hyperparameters $\tau_\epsilon, \kappa_\epsilon$. More specifically, the posterior density of σ^2_α is

$$\begin{aligned} p(\sigma^2_\alpha \mid Y, \sigma^2_\epsilon) \propto\ & (\sigma^2_\alpha + \kappa_\epsilon \sigma^2_\epsilon)^{-u_\alpha - 1} (\varsigma^2_{\alpha\epsilon} + \frac{\sigma^2_\epsilon}{n_J})^{-n_I/2} \\ & \times \exp\left(-\frac{v_\alpha}{\sigma^2_\alpha + \kappa_\epsilon \sigma^2_\epsilon} - \frac{1}{2}\frac{1}{\varsigma^2_{\alpha\epsilon} + \frac{\sigma^2_\epsilon}{n_J}} \sum_i (\widehat{\alpha}_i - \alpha_0)^2\right), \end{aligned}$$

where $\varsigma^2_{\alpha\epsilon} = (\frac{\tau_\alpha}{\sigma^2_\alpha} + \frac{\tau_\epsilon}{\sigma^2_\epsilon})^{-1}$, which can be seen as a type of shifted inverse-gamma distribution. However, rather than normalizing the posterior density so that it is proper when constrained to non-negative values, it is often more informative to consider a mass point at zero; allowing for the hypothesis (7) to be tested (see Section 4).

Each individual α_i is however normally distributed, with posterior density

$$p(\alpha_i \mid Y, \sigma_\epsilon^2, \sigma_\alpha^2) \propto Q_\alpha^{1/2} \exp\left(-\frac{Q_\alpha}{2}\left[\alpha_i - Q_\alpha^{-1}\left(\frac{1}{\varsigma_{\alpha\epsilon}^2}\alpha_0 + \frac{n_J}{\sigma_\epsilon^2}\widehat{\alpha}_i\right)\right]^2\right),$$

where $Q_\alpha = \frac{1}{\varsigma_{\alpha\epsilon}^2} + \frac{n_J}{\sigma_\epsilon^2} = \frac{\tau_\alpha}{\sigma_\alpha^2} + \frac{\tau_\epsilon + n_J}{\sigma_\epsilon^2}$.

3.1.2 Conjugate Prior

Setting $\tau_\epsilon = 0, \kappa_\epsilon = \frac{\tau_\alpha}{n_J}$, for the prior and likelihood

$$(\alpha_i, \sigma_\alpha^2, \sigma_\epsilon^2) \sim \ \mathrm{N}\Gamma^{-1}\Gamma^{-1}(\alpha_0,\ \tau_\alpha,\ \tau_\epsilon = 0,\ \kappa_\epsilon = \frac{\tau_\alpha}{n_J},\ u_\alpha,\ v_\alpha,\ u_\epsilon,\ v_\epsilon), \tag{20}$$

$$Y_{ij} \mid \alpha_i, \sigma_\epsilon^2 \sim \ \mathrm{N}(\alpha_i,\ \sigma_\epsilon^2), \tag{21}$$

gives way to the posterior distribution

$$(\alpha_i, \sigma_\alpha^2, \sigma_\epsilon^2) \mid Y \sim \ \mathrm{N}\Gamma^{-1}\Gamma^{-1}\left(\left[\frac{\tau_\alpha}{\sigma_\alpha^2} + \frac{n_J}{\sigma_\epsilon^2}\right]^{-1}\left[\frac{\tau_\alpha}{\sigma_\alpha^2}\alpha_0 + \frac{n_J}{\sigma_\epsilon^2}\widehat{\alpha}_i\right],\ \tau_\alpha,\ n_J,\ \frac{\tau_\alpha}{n_J},\ u_\alpha + \frac{n_I}{2},\right.$$
$$\left. v_\alpha + \frac{\tau_\alpha}{2}\sum_i(\widehat{\alpha}_i - \alpha_0)^2,\ u_\epsilon + \frac{n - n_I}{2},\ v_\epsilon + \frac{1}{2}\sum_{ij}(y_{ij} - \widehat{\alpha}_i)^2\right), \tag{22}$$

thus maintaining conjugacy. Particularly beneficial is that with this factorization there is no need for MCMC sampling. Rather, posterior draws can be taken immediately without burn-in nor thinning. A single realization from the joint posterior is found by sampling from $p(\sigma_\epsilon^2|Y)$, $p(\sigma_\alpha^2|Y, \sigma_\epsilon^2)$, and then $p(\alpha_i|Y, \sigma_\epsilon^2, \sigma_\alpha^2)$ using (19), (18), (17) with the parameters updated using the observations.

Selecting appropriate values hyperparameters can then be done as follows. For invariant priors of variance parameters, hyperparameters are $u_\alpha = v_\alpha = u_\epsilon = v_\epsilon = 0$, which may be considered as limiting forms of proper inverse-gamma distributions. To maintain conjugacy $\tau_\epsilon = 0, \kappa_\epsilon = \frac{\tau_\alpha}{n_J}$ is used. Practical values of τ_α and α_0 may then be found using an empirical Bayes approach and yield $\tau_\alpha = 1$ and $\alpha_0 = n_I^{-1}\sum_i \widehat{\alpha}_i$, i.e. the overall mean.

For more general models, with a mean term, additional factors, interactions, etc., it is possible to consider several such normal–inverse-gamma–inverse-gamma distributions, where the single inverse-gamma distribution of the errors, σ_ϵ^2, is common to all. One may then use a prior distribution for the factor levels under a linear constraint so that the

posterior distributions can also be factored similarly. This allows not only for conjugacy, but also facilitates computation in a way that even for models with many factors, samples from the posterior can efficiently be drawn without the need for MCMC. This is seen in Geinitz *et al.* (2012).

4 Examples

4.1 Rails Data

For illustration of the various analysis variance methods consider the balanced one-way design for data consisting of six railway rails (Devore, 2000; Pinheiro and Bates, 2000). Each rail has been measured three times for the amount of time that it takes a certain type of ultrasonic wave to travel the length of the rail. The objective of any initial analysis is most likely to investigate the (a) variation due to measurement error and (b) variation due to the rails themselves in terms of both statistical and practical significance. Additionally, one may be interested predicting travel time for a future measurement. This can be considered for either (c) one of these rails as well as (d) a future rail that has not yet been seen. For this one-way analysis we consider the simple cell-means model

$$Y_{ij} = \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, 6, \quad j = 1, 2, 3. \tag{23}$$

4.1.1 Conventional Methods

For this one-way model error terms ϵ_{ij} are assumed to be iid $\mathrm{N}(0, \sigma_\epsilon^2)$, and group terms α_i as unknown constants, so that the observations are

$$Y_{ij} \mid \alpha_i, \sigma_\epsilon^2 \sim \mathrm{N}(\alpha_i,\, \sigma_\epsilon^2). \tag{24}$$

The model assumes that the six rails that have been observed are the only rails that are of interest. This is a fixed effects model, which is to say that the population of rails has been exhausted by the sample.

Questions (a) and (b) can be reasonably addressed using this model. An estimate of the variation due to the measurement error uses the MSE, which is seen directly in the standard ANOVA table. The variation due to the rails is formally tested using the

hypothesis of (5). The corresponding p-value for this test is also reported in Table 3. There is, however, no formal technique for testing the practical relevance of the variation due to the rails.

Table 3: One-way ANOVA of Railway Rails

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Rail	5	9310.50	1862.10	115.18	0.000000
Residuals	12	194.00	16.17		

Question (c) could be answered by looking at the standard error for the estimate $\widehat{\alpha}_i$, to obtain an estimate of the expected travel time. Question (d) can, however, not be answered because of the assumed fixed effect. To address this question the rails must be considered as a random effect, i.e. assumed to come from a greater population of rails.

The random effects model is able to addresses (a) and (b) similarly, although the interpretation of (b) is that the observed rails are only a subset of a much larger (possibly infinite) population of possible rails. Hence, only statistical significance is again considered and is done using the hypothesis of (7), which as discussed in Section 2.1.2, uses the same F-statistic as for the fixed effects model.

Although the ratio of $\frac{\sigma_\alpha^2}{\sigma_\epsilon^2}$ may be estimated as a means to gauge practical relevance of variation due to the rails, this is often an additional step that is not reported by default for random effects procedures by most software packages. A more informative summary is often to identify a confidence region of the variance components, $\sigma_\alpha^2, \sigma_\epsilon^2$, as seen using the likelihood function in Figure 2.

4.1.2 Comprehensive ANOVA

Analogous to the ANOVA summary provided by Table 3, but including both finite and superpopulation variances, Table 4 presents a clearer view on the practical significance of the rails. Figure 3 similarly summarizes the analysis. From the graphical plot statistical significance is suggested by the fact that the intervals do not extend to 0. In contrast to this, the next example illustrates how negative posterior support can be used to test

statistical significance.

Table 4: Bayesian ANOVA Table: Posterior distributions are used to obtain two estimates of the variability. Quantiles provide an assessment of the uncertainty in these estimates. The probability, e.g. $\Pr(\sigma_\alpha > \sigma_\epsilon)$, provide a relative comparison of each variance parameter to the measurement variability.

	Parameter	Mean	$Q_{0.5}$	$(Q_{0.025}, Q_{0.975})$	$\Pr(> \sigma_\epsilon)$
Rails	(finite) s_α	24.69	24.71	$(22.46, 26.83)$	1.000
	(super) σ_α	25.96	23.89	$(14.55, 49.20)$	1.000
Errors	σ_ϵ	4.27	4.10	$(2.87, 6.58)$	–

4.2 Simulated Data

In the following example a comparison of practical and statistical significance is illustrated using both classical ANOVA as well as the more comprehensive Bayesian ANOVA. Data of the form $Y_{ij} = \alpha_i + \epsilon_{ij}$ is generated where $\alpha_i \sim \mathrm{N}(0, \sigma_\alpha^2)$ and $\epsilon_{ij} \sim \mathrm{N}(0, \sigma_\epsilon^2 = 1)$, for $i = 1, \ldots, n_I,\ j = 1, \ldots, n_J$ for a total of $n = n_I \cdot n_J$ observed values. This is done under two distinct cases

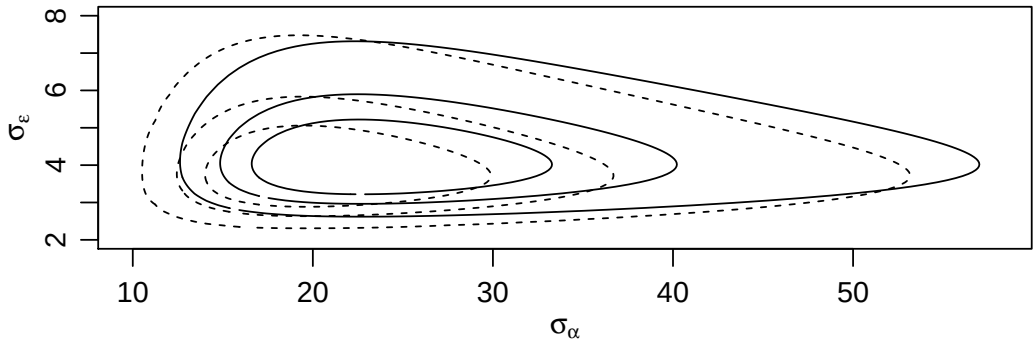

Figure 2: Comparison of confidence regions for superpopulation standard deviations based on χ^2 approximation of relative log-likelihood (solid) and using the highest-posterior-density (dotted). Contours correspond to confidence levels $0.50, 0.75$, and 0.95 (small to large).

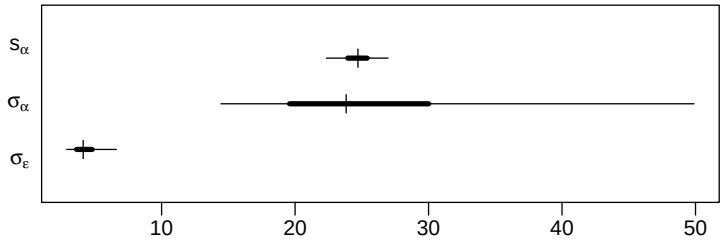

Figure 3: Graphical summary of posterior quantiles for variance component (super and finite population) parameters.

$$\text{Case A:} \quad \sigma_\alpha^2 = \tfrac{1}{2}, \quad n_J = 6,$$
$$\text{Case B:} \quad \sigma_\alpha^2 = 2, \quad n_J = 2,$$

and with $n_I = 5$ for both. Using the conventional ANOVA method (Table 5) there is not any discernible differences between the two datasets. Statistical significance is approximately equivalent because of the balance of statistical power and the difference in the variance components σ_α^2 and σ_ϵ^2.

By assembling a mass point at zero whenever the posterior has support for negative values it possible to use the probability $p(\sigma_\alpha^2 = 0|Y)$ to test the hypothesis of (7). Interestingly, this posterior probability is 0.0263 for case A and 0.0258 for case B, values which are comparable to their corresponding p-values in Table 5. This is also noted by the end-points corresponding to the 0.025 level of uncertainty for the intervals shown in Figure 4. Note that because this is restricted to a simple hypothesis the usual challenges of finding a posterior predictive p-value (Bayarri and Berger, 2004) are avoided.

A more informative and comprehensive summary of the data is provided by the Bayesian ANOVA table (Table 6). This provides not only estimates of the variance components, but also an indication of the practical significance of the factor α when observed with error ϵ.

5 Discussion

The major contribution of this paper can be seen as ideological in nature, in that the standard method of analysis of variance is treated as a useful procedure to practitioners.

Table 5: Classical ANOVA table to summarize the decomposition of variance. Case A (left), with $n_I = 5, n_J = 6, \sigma_\alpha^2 = \frac{1}{2}$, represents low practical significance of factor α. Case B (right), with $n_I = 5, n_J = 2, \sigma_\alpha^2 = 2$, represents strong practical significance. Despite practical differences between the two cases, p-values are nearly equal.

	Df	Sum Sq	Mean Sq	F val	Pr(>F)
α	5	9.70	1.94	3.08	0.0267
ϵ	25	15.75	0.63		

	Df	Sum Sq	Mean Sq	F val	Pr(>F)
α	5	27.69	5.54	7.31	0.0239
ϵ	5	3.79	0.76		

Table 6: Bayesian ANOVA tables to summarize the variance decomposition. Case A (left) and Case B (right) illustrate a situation in which factor α has low, or high practical significance.

	Par	Mean	$Q_{0.5}$	$(Q_{0.025}, Q_{0.975})$	$\Pr(> \sigma_\epsilon)$
α	s_α	0.48	0.49	$(0.01, 0.83)$	0.07
	σ_α	0.57	0.51	$(0.00, 1.37)$	0.19
ϵ	σ_ϵ	0.82	0.80	$(0.62, 1.09)$	–

	Par	Mean	$Q_{0.5}$	$(Q_{0.025}, Q_{0.975})$	$\Pr(> \sigma_\epsilon)$
α	s_α	1.57	1.63	$(0.08, 2.33)$	0.83
	σ_α	1.77	1.61	$(0.00, 3.97)$	0.83
ϵ	σ_ϵ	1.03	0.93	$(0.55, 2.14)$	–

As a result, although rigorous treatment is given, the discussion has been restricted to relatively simple designs. Extending from a one-way balanced ANOVA to many factors can be considered as a trivial step from here. However, much more work is needed in order for the method to be able to be widely accepted.

Specific Bayesian aspects to be further developed are the precise usage of p-values and of the prior specification. Bayarri and Berger (2004) have used a posterior predictive p-value that has been marginalized over the unknown parameter, which allows for composite hypothesis testing. Brown and Walker (2012) have shown that a nonparametric prior can be used to specify a more general prior, which they deem as a compromise between informative and noninformative. From a broader point of view, the issues of unbalanced designs, non-orthogonal predictors, and generalized linear models are also necessary for a thorough development of the statistical method.

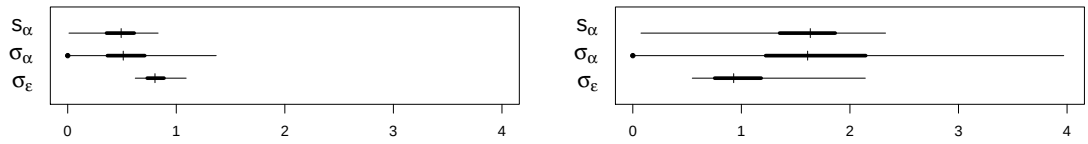

Figure 4: Posterior uncertainty intervals of variance parameters shown graphically for both cases. Thick line segments correspond to 50% uncertainty and thin line segments to 95%. Vertical marks denote the posterior median. A point at the end of an interval denotes evidence that the variance component is zero and is analogous to the hypothesis of a null effect.

References

Bates, D. and DebRoy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, **91**, 1–17.

Bayarri, M. J. and Berger, J. O. (2004). The interplay of bayesian and frequentist analysis. *Statistical Science*, **19**, 58–80.

Bernardo, J. M. and Smith, A. (2000). *Bayesian Theory (Wiley Series in Probability and Statistics)*. John Wiley and Sons Ltd.

Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*. Wiley-Interscience.

Brown, P. J. and Walker, S. G. (2012). Bayesian priors from loss matching. *International Statistical Review*, **80**, 60–82.

Cochran, W. G. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society*, **30**, 178–191.

Cox, D. and Solomon, P. J. (2003). *Components of Variance*, volume 97 of *Monographs on statistics and applied probability*. Chapman and Hall.

Dawid, A., Stone, M., and Zidek, J. V. (1973). Marginalization paradoxes in bayesian and structural inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, , 189–233.

Devore, J. (2000). *Probability and Statistics for Engineering and the Sciences*. Brooks/Cole, Thomson Learning.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics*, **3**, pp. 1–21.

Everson, P. J. and Morris, C. N. (2000). Inference for multivariate normal hierarchical models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **62**, pp. 399–412.

Fidler, F., Geoff, C., Mark, B., and Neil, T. (2004). Statistical reform in medicine, psychology and ecology. *Journal of Socio-Economics*, **33**, 615–630.

Fisher, R. (1925). *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1 edition.

Gardner, M. and Altman, D. (1986). Confidence intervals rather than p values: estimation rather than hypothesis testing. *British medical journal (Clinical research ed.)*, **292**, 746.

Geinitz, S., Furrer, R., and Sain, S. R. (2012). Multivariate analysis of global climate projections via rank deficient Bayesian ANOVA. *Journal of the Royal Statistical Society: Series C (under revision for resubmission)*.

Gelman, A. (2005). Analysis of variance: Why it is more important than ever. *Annals of Statistics*, **33**, 1–31.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 1 edition.

Goldstein, H. (1995). *Multilevel Statistical Models*. Kendall's Library of Statistics. E. Arnold.

Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology*, **29**, 158–167.

Hald, A. (1998). *A history of mathematical statistics from 1750 to 1930*. Wiley Series in Probability and Statistics. Wiley.

Hoaglin, D., Mosteller, F., and Tukey, J. (1991). *Fundamentals of Exploratory Analysis of Variance*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Huber, P. (2011). *Data Analysis: What Can Be Learned From the Past 50 Years*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Public Library of Science Medicine*, **2**, e124.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **186**, pp. 453–461.

Kreft, I., De Leeuw, J., and de Leeuw, J. (1998). *Introducing multilevel modeling*. Sage London.

Leinonen, T., O'Hara, R. B., Cano, J. M., and Merilä, J. (2008). Comparative studies of quantitative trait and neutral marker divergence: a meta-analysis. *J. Evolutionary Bio.*, **21**, 1–17.

Lencina, V. B., Singer, J. M., and Stanek, E. J. (2005). Much ado about nothing: the mixed models controversy revisited. *International Statistical Review*, **73**, 9–20.

Lindsay, B. G., Kettenring, J., and Siegmund, D. (2004). A Report on the Future of Statistics. *Statistical Science*, **19**, 387–413.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.

Nakagawa, S. and Cuthill, I. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, **82**, 591–605.

Nelder, J. A. (1999). From statistics to statistical science. *J. R. Statist. Soc. D*, **48**, 257–269.

Nelder, J. A. (2008). What is the mixed-models controversy? *International Statistical Review*, **76**, 134–135.

Pinheiro, J. and Bates, D. (2000). *Mixed-Effects Models in S and S-Plus*. Statistics and Computing. Springer.

Pinheiro, J., Bates, D., DebRoy, S., and Sarkar, D. (2006). *nlme: Linear and nonlinear mixed effects models*.

Qian, S. S. and Shen, Z. (2007). Ecological applications of multilevel analysis of variance. *Ecology*, **88**, 2489–2495.

Rao, P. (1997). *Variance components: mixed models, methodologies and applications*, volume 78. Chapman & Hall/CRC.

Sain, S. R., Nychka, D. W., and Mearns, L. (2011). Functional ANOVA and regional climate experiments: a statistical analysis of dynamic downscaling. *Environmetrics*, **22**, 700–711.

Savage, I. R. (1957). Nonparametric Statistics. *Journal of the American Statistical Association*, **52**, 331–344.

Searle, S., Casella, G., and McCulloch, C. (1992). *Variance components*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Snijders, T. and Bosker, R. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publications Limited.

Von Korff, M., Koepsell, T., Curry, S., and Diehr, P. (1992). Multi-level analysis in epidemiologic research on health behaviors and outcomes. *American Journal of Epidemiology*, **135**, 1077–1082.

Voss, D. T. (1999). Resolving the mixed models controversy. *American Statistician*, **53**, 352–356.

Yoccoz, N. (1991). Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America*, **72**, 106–111.

Paper II

Multivariate analysis of global climate projections via rank deficient Bayesian ANOVA

Steven Geinitz & Reinhard Furrer & Stephan R. Sain

Paper to be submitted.

Multivariate analysis of global climate projections via rank deficient Bayesian ANOVA

Steven Geinitz	Reinhard Furrer	Stephan R. Sain
geinitz@math.uzh.ch	*furrer@math.uzh.ch*	*ssain@ucar.edu*
University of Zurich	University of Zurich	National Center for Atmospheric Research
Zurich, Switzerland	Zurich, Switzerland	Boulder, CO, U.S.A.

Abstract Simulated climate projections are subject to distinct sources of variation. Two primary sources are the numerical climate model and the presumed emissions scenario. Identifying appropriate statistical parameters to represent the sources of variability, utilizing computationally efficient procedures, and quantifying multivariate sources of variation are challenges that any general method must take into consideration. A hierarchical Bayesian analysis of variance method is presented here, which allows for variance parameters to be assessed simultaneously in both a fixed effect and random effect context. By specifying improper prior distributions with rank deficient covariances factorization of the full joint posterior can be carried out, which obviates the need for Markov chain Monte Carlo methods. For each multivariate source of variability, an assessment of its magnitude, the uncertainty in the assessment, as well as a relative comparison across all sources of variation is obtained. The method is illustrated with simulated data and with climate projections made available as part of the coupled model intercomparison project (CMIP3), in which global decadal temperature and precipitation variability due to general circulation climate models and to emissions scenarios is examined.

Keywords: hierarchical Bayes; constraints; finite-population; super-population; uncertainty quantification; variance components;

1 Introduction

Simulated climate projections have the potential to profoundly impact current economical, political, and social policy decisions being made. Thus, the ability to quantify the distinct sources of uncertainty within these types of experiments is necessary. Projections are made by many research facilities around the globe, each using one or more atmosphere-ocean general circulation model (GCM) to discretize and numerically solve the fluid dynamical systems representing large scale climate processes. In addition, the system is constrained to forcings that represent, primarily, human-driven impacts on the composition of gases in the atmosphere. Distinct forcings have been identified with different emission paths, or scenarios (SCN), and have been formulated in the Special Report on Emissions Scenarios (Nakićenović and Swart, 2000). The climate projections discussed in this article have been used for the fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC) (Meehl *et al.*, 2007). Two primary sources of variation in overall climate projection variability that must be quantified are then GCM and SCN, each of which is a factor consisting of a set of factor levels.

It should be stressed that the data analysed herein consist of simulated climate projections and not of directly observed or recorded climate measurements. However, from a general statistical point of view, the data can be considered as *observations* originating from climate simulation experiments, hence, they are regarded as having been *observed*.

In conventional analysis of variance terminology, a set of factor levels may be assumed as (exclusively) either fixed or random, and a statistical model of that name is used. The fixed effects model focuses on the factor levels present in the observed data or experiment, while the random effects model considers all potential factor levels, even those that have not been observed. Using a hierarchical Bayesian model we examine each of these cases simultaneously by examining variability of the factor levels in the observed data, as well as variability of factor levels that have not yet been observed. This distinction is made by using the terms finite-population variance and super-population variance, respectively. In the context of the application presented here this means that the variance of the existing GCM factor levels can be considered and would be denoted as the finite-population variance. Whereas the variance of unobserved GCM factor levels, i.e. those that do not yet exist, correspond to the super-population variance.

For each type we assess the magnitude of each source of variation, compare all sources

of variation, and provide an estimate of the uncertainty thereof. A similar strategy has been used by Kaufman and Sain (2010), in which uncertainty from general circulation model and from regional climate models is examined, although uncertainty of super-population variability is not explicitly investigated. Yip *et al.* (2011) use an ANOVA approach as well when considering uncertainty of average global temperatures, but also consider only the finite-population case. Results of the analysis herein suggest that variation in climate projections is due primarily to GCM. Although variation due to SCN is suggested to be smaller, the uncertainty in this assessment remains high. Additionally, we examine the sources of variability present when considering relative decadal climate change up to 2100 when uncertainty of existing factor levels, as well new factor levels, is considered.

The hierarchical Bayesian model utilized herein can be seen as a type of multilevel model, which itself is defined as a linear, or generalized linear, regression model in which the regression coefficients are given a probability model, i.e. varying coefficients as well as a model for these varying coefficients (Gelman and Hill, 2006). Multilevel models have become more widely used because of their ability to examine data structured in a hierarchical, or nested form. An additional advantage is in being able to analyse the statistical parameters related to variability in the data for all potential sources of variation. These types of statistical models are seeing usage in diverse fields such as ecology (Qian and Shen, 2007), genetics (Leinonen *et al.*, 2008), and climate (Sain *et al.*, 2011), to name a few. Two potential limitations still exist in how such types of analyses are typically carried out. First, computationally intensive MCMC (Hastings, 1970) procedures are used almost exclusively to obtain posterior distributions of variance parameters. Second, to facilitate convergence, additional parameters are introduced so that dependence among the effects within a single source of variation may be reduced (Gelman, 2005). Such problems can be compounded further when dealing with climate model data and, more generally, in multivariate cases.

This paper confronts these challenges by first addressing dependencies within a single source of variation, which obviates the need for MCMC for a wide range of models. The strategy for doing so is based on the use of improper joint prior distributions with rank deficient covariances. This is related to identifiability in classical models and has been an ongoing topic of discussion by many, spanning from Nelder (1977) to Nelder (2008), and

references therein. With the rank deficiencies, it is then possible to factor the full joint posterior distribution in an efficient manner.

In Section 2 we review the univariate multilevel ANOVA formulation of Gelman (2005). Section 3 rigorously focuses on the details of our multivariate rank deficient Bayesian approach, which incorporates parameter constraints into priors, and discusses computation for general cases. Section 4 provides a demonstration of the method through a simulation example. Section 5 applies the method to the climatological data thus far described. In Section 6 we discuss extensions and further computational benefits that are possible when the method is applied to high-dimensional problems.

2 Background

Analysis of variance is widely accepted as a means of partitioning variability in a manner that allows it to be attributed to various factors. An important initial step in the analysis is considering each factor of the model to be fixed or random. Simple definition of the terms, fixed and random, is, however, not straightforward. Searle *et al.* (1992), Kreft *et al.* (1998), and Snijders and Bosker (1999) each offer different definitions. Gelman (2005) reviews these and others.

Assuming that one definition has been accepted, it is still not clear in practice whether a set of parameters in a statistical model should be specified as fixed or random. For example, in a statistical model with a factor α and factor levels $\alpha_1, \ldots, \alpha_{n_\alpha}$, it is possible that interest lies in both the estimates of the individual α_i and in its super-population variance, σ^2_α. Searle *et al.* (1992) pragmatically recommend the random effects model to be used whenever the additional assumption is deemed reasonable, and fixed otherwise.

Further problems arise when more terms are present in the statistical model. For $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, with α fixed and β random and interaction term $(\alpha\beta)$, it is not entirely clear how the identifiability constraints should influence the distribution of the random components β_j. Herein lies the so-called mixed models controversy (Voss, 1999). This issue leads to disagreements of which sums of squares quantities should be used in the test of a significance for the main random effect. Discussions in the literature on this issue have prolonged. Nelder (1977) has reiterated his views on parameter constraints in a series of papers (Nelder, 1994; Nelder *et al.*, 1998; Nelder, 2008), and has led authors

to reconsider their previous positions, as with (Voss, 1999, 2000) and (Lencina *et al.*, 2005; Lencina and Singer, 2006). Due to the continued debate, one may assume that a universally agreed upon solution has not yet been found.

A preliminary move needed to address the controversy is to indiscriminately consider all components in a model as random. This initial step may be seen as a fundamental contribution of the hierarchical regression approach to ANOVA employed by Gelman (2005), as it eliminates the need to define terms as fixed or random and facilitates comparison across all sources of variability. Factors are then considered more generally, so that they apply to each potential source of variability in the model, e.g. overall mean, factors, nested terms, interactions, etc. These are referred to as generalised factors, but simply factors herein. The nature of the variability from a factor is then further distinguished. The distinction traditionally made by fixed and random effects is instead addressed by considering a factor's *finite-* and *super*-population variances. Following Gelman and Hill (2006), the finite-population variance of a factor is defined as the variability of the set of factor levels in the existing data, while the super-population variance corresponds to the uncertainty of a new factor level that has not yet been observed.

2.1 Model Parametrization

Following the notation of Gelman (2005), observations $Y_i, i = 1, \ldots, n$, are stated in terms of the additive decomposition

$$Y_i = \sum_{b=0}^{B} \beta_{j_i^b}^{(b)}. \tag{1}$$

An individual factor is referenced by $b = 0, \ldots, B$, with each consisting of n_b distinct factor levels. Factor indices $b = 0$ and $b = B$ will often correspond to an overall mean, μ, and to measurement errors, ϵ_i, respectively, so that $n_0 = 1$, $n_B = n$. In practice the number of factors B is reasonable so that a conventional ANOVA notation can be used entirely. However, for discussion of the general case the additional level of sub and super scripting is needed. That is, for the balanced model

$$Y_{rst} = \mu + \alpha_r + \beta_s + (\alpha\beta)_{rs} + \epsilon_{rst}, \tag{2}$$

with $r = 1, \ldots, n_r$, $s = 1, \ldots, n_s$, and $t = 1, \ldots, n_t$. Factor α then corresponds to $b = 1$, factor β to $b = 2$, etc., so that this model is stated more generally as

$$Y_i = \beta^{(0)} + \beta^{(1)}_{j_i^1} + \beta^{(2)}_{j_i^2} + \beta^{(3)}_{j_i^3} + \beta^{(4)}_{j_i^4}, \tag{3}$$

with i ranging from 1 to $n = n_r n_s n_t$. An individual factor level is referenced equivalently by either α_r or $\beta^{(1)}_1$. The subscripts $j^b_1, \ldots, j^b_n$ replicate the factor levels appropriately so that exactly one level from each factor is associated with exactly one observation. Specifically, the set of subscript indices for the factor $b = 1$, i.e. α, is

$$\begin{aligned} &\{j^1_1, \ldots, j^1_{n_s n_t}, j^1_{n_s n_t+1}, \ldots, j^1_{2n_s n_t}, \ldots\ldots\ldots\ldots, j^1_{(n_r-1)n_s n_t+1}, \ldots, j^1_n\} \\ &= \{1, \ldots, 1, \quad 2, \quad \ldots, \quad 2, \quad \ldots\ldots\ldots\ldots, \quad n_r, \quad \ldots, n_r\}, \end{aligned}$$

which is used to replicate the distinct factor levels to match the indices for the full set of observations. The set of distinct levels for this factor are denoted as $\beta^{(1)} = \{\beta^{(1)}_1, \ldots, \beta^{(1)}_{n_r}\}$ using model (3), and is equivalent to $\alpha = \{\alpha_1, \ldots, \alpha_{n_r}\}$ when using (2).

Similarly, a regression formulation may also be used when considering covariates. The extension of (1) is then

$$Y_i = \sum_{b=0}^{B} \sum_{j=1}^{n_b} x^{(b)}_{i,j} \beta^{(b)}_{j^b}, \tag{4}$$

with $\beta^{(b)}_{j^b}$ denoting a factor's distinct levels and $x_{i,j}$ denoting explanatory variables. The regression formulation could be used for the additive decomposition, with explanatory variables set to either 0 or 1.

Given a factor, and an assumed distribution on model errors, a conventional fixed effects analysis often corresponds to the test of the null hypothesis $H_0 : \beta^{(b)}_j = 0$ for $j = 1, \ldots, n_b$. While for a random effects model, assuming the n_b levels of each factor b to be modelled as Gaussian,

$$\beta^{(b)}_j \mid \sigma^2_b \overset{iid}{\sim} \mathrm{N}(0,\ \sigma^2_b), \qquad j = 1, \ldots, n_b, \tag{5}$$

a test for significant variation of a factor would be $H_0 : \sigma^2_b = 0$. Alternatively, the method identifies two representations of variation for a given factor. The super-population variance, σ^2_b, corresponds to the variance of all potential, possibly infinitely many, factor levels. The finite-population variance represents variability of the specific set of factor levels that have been realized. Super and finite-population variances can be roughly related to the random effect variance component estimate, and the fixed effect within-group sum of squares, respectively. Consider the set of levels of factor b as a vector, $\boldsymbol{\beta}^{(b)} =$

$(\beta_1^{(b)}, \ldots, \beta_{n_b}^{(b)})^T$, with c_b constraints. Degrees of freedom for the factor is then $\nu_b = n_b - c_b$, and the finite-population variance s_b^2 is $s_b^2 = \frac{1}{\nu_b}\boldsymbol{\beta}^{(b)^T}\left(\mathbf{I}_{n_b} - \mathbf{C}_b^T(\mathbf{C}_b\mathbf{C}_b^T)^{-1}\mathbf{C}_b\right)\boldsymbol{\beta}^{(b)}$, where $\mathbf{I}_{n_b}$ is the $n_b \times n_b$ identity and $\mathbf{C}_b$ is a $c_b \times n_b$ constraint matrix such that $\mathbf{C}_b\boldsymbol{\beta}^{(b)} = \mathbf{0}$. Variance component estimation is made by decomposing the estimated variance of the factor levels, $V_b = \mathrm{Var}(\widehat{\beta}^{(b)})$, into the sum of the super-population variation, σ_b^2, plus the variability of the factor level estimates, $V_{\text{b:estimation}}$, based on the law of total variance

$$V_b = \mathrm{Var}(\widehat{\beta}^{(b)})) = \mathrm{Var}\{\mathrm{E}(\widehat{\beta}^{(b)} \mid \beta^{(b)})\} + \mathrm{E}\{\mathrm{Var}(\widehat{\beta}^{(b)} \mid \beta^{(b)}) = \sigma_b^2 + V_{\text{b:estimation}}.$$

A natural estimate of the super-population variance is then the method-of-moments estimator

$$\widehat{\sigma}_b^2 = \widehat{V}_b - \widehat{V}_{\text{b:estimation}}, \tag{6}$$

where $\widehat{V}_b = \frac{1}{\nu_b}\sum_{j=1}^{n_b}(\widehat{\beta}_j^{(b)})^2$, and $\widehat{V}_{\text{b:estimation}} = \sum_{k\in I(b)}\frac{n_b}{n_k}\widehat{\sigma}_k^2$ includes super-population variances from other factors that contribute to the variability of the factor level estimates, indicated by the set $I(b)$. This is analogous to a generalization of the sums of squares decomposition. For the simple balanced one-way random effects model, $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, with $\mathrm{Var}(Y_{ij}) = \sigma_\alpha^2 + \sigma_\epsilon^2$, $n = n_I n_J$, and sums of squares decomposition $\mathrm{SST} = \mathrm{SSA} + \mathrm{SSE}$; the mean squared error terms of which are MSA, MSE and have expectations $n_J\sigma_\alpha^2 + \sigma_\epsilon^2, \sigma_\epsilon^2$, respectively. Thus, the estimator of σ_α^2 is distributed as a linear combination of chi-square random variables, $\chi_{\nu_A}^2 - \frac{1}{n_J}\chi_{\nu_E}^2$ where ν_A, ν_E are the appropriate degrees of freedom. At a minimum, $\widehat{V}_{\text{b:estimation}}$ will include $\widehat{\sigma}_B^2$, the estimated error variance. For models in which all factors are orthogonal to one another, $\widehat{V}_{\text{b:estimation}} = \frac{n_b}{n}\widehat{\sigma}_B^2$, $b = 0, \ldots, B-1$.

2.2 Confirmatory Steps

In regards to more inferential procedures, either a frequentist or Bayesian direction can be taken. In the frequentist case, an inverse-chi-square distribution, χ_ν^{-2}, is employed to assess uncertainty in the super-population variance σ_b^2; since the sample variance divided by $\nu_b\sigma_b^2$ is distributed as a chi-square random variable with ν_b degrees of freedom. For factors with $\sum_{k\in I(b)}\frac{n_b}{n_k}\widehat{\sigma}_k^2$, including more than only the error variance $\widehat{\sigma}_B^2$, then a linear combination of inverse-chi-square distributions as in (6), $\sum_i m_i\chi_{\nu_i}^{-2}$, is required. As Gelman (2005) states, these linear combinations may be dealt with directly, although simulation is often more straightforward. The simulation, described therein, is carried out

by: 1) Obtain R simulated raw variances, V_b for each of the B factors in the model with a random variable that is proportional to χ^{-2} and corresponding degrees of freedom; 2) Calculate super-population variances using $\widehat{\sigma}_b^2 = \max(0, \widehat{V}_b - \widehat{V}_{\text{b:estimation}})$; 3) Simulate factor levels using newly generated super-population variances; 4) Calculate sample variances of each set of factor levels. This informal procedure then yields a (posterior) sample of super-population variances, factor levels, and finite-population variances, corresponding to the final three steps. While this ad-hoc method provides useful insight, multilevel models have typically come to rely on MCMC methods, often Gibbs samplers, e.g. `WinBUGS` (Lunn *et al.*, 2000). We follow the strategy as well. However, by using an explicitly Bayesian formulation, and applying constraints to the prior distribution of the set of factor levels, it is often possible to eliminate computationally demanding MCMC procedures, as seen in the following section.

3 Multivariate Analysis of Variance

Using an explicitly Bayesian model, we now derive a general analysis of variance methodology that achieves the same intuitive results on all sources of variation as in Section 2. Thus, a joint prior distribution is specified for the set of levels in a factor. By applying a constraint on this joint distribution, an improper joint prior distribution results. This avoids the issues of constraints that arise, as in the mixed model controversy, and also allows for a computationally convenient factorization of the full posterior distribution. In the case of balanced designs the joint posterior is factored into a factor's super-population covariance posterior and each of its distinct factor level posteriors. As a result it is possible to sample directly from the joint posterior in an independent manner, thereby obviating the need for MCMC. Since more general, higher dimensional cases are considered, the multivariate form of the model is given. This is now examined in greater depth.

3.1 Multivariate Model Parameterization

Consider d-dimensional multivariate observations such that factor levels are vectors and a factor's variance is a covariance matrix. Namely, (1)–(5) now contain vectors $\mathbf{Y}_i$, $\boldsymbol{\beta}_{j^b}^{(b)}$, matrices $\mathbf{X}_{i,j}$, and covariance matrices are $\boldsymbol{\Sigma}_b$, all of appropriate dimension. The multivariate analogue of (1) with Gaussian errors is

$$\mathbf{Y}_i \mid \{\boldsymbol{\beta}_{j_i^b}^{(b)}\}_{b=0}^{B-1}, \boldsymbol{\Sigma}_\epsilon \sim \mathrm{N}_d\left(\sum_{b=0}^{B-1} \boldsymbol{\beta}_{j_i^b}^{(b)},\ \boldsymbol{\Sigma}_\epsilon\right), \qquad i = 1, \ldots, n, \tag{7}$$

where $\{\boldsymbol{\beta}_{j^b}^{(b)}\}_{b=0}^{B-1}$ is the set of factors over the given indices. The remainder of the specification is given by

$$\boldsymbol{\beta}_j^{(b)} \mid \boldsymbol{\Sigma}_b \sim \mathrm{N}_d(\boldsymbol{\beta}_0^{(b)},\ \boldsymbol{\Sigma}_b), \qquad b = 0, \ldots, B, \quad j = 1, \ldots, n_b, \tag{8}$$

$$\mathbf{U}_b = \boldsymbol{\Sigma}_b + \frac{n_b}{n}\boldsymbol{\Sigma}_\epsilon \mid \boldsymbol{\Sigma}_\epsilon \sim W^{-1}(\boldsymbol{\Psi}_b,\ \kappa_b), \qquad b = 0, \ldots, B-1, \tag{9}$$

$$\boldsymbol{\Sigma}_\epsilon \sim W^{-1}(\boldsymbol{\Psi}_\epsilon,\ \kappa_\epsilon), \tag{10}$$

where the inverse-Wishart distribution is denoted by W^{-1}. Factor indices $b = 0$ and $b = B$ respectively correspond to the intercept, or overall mean term, $\boldsymbol{\mu}$, and to errors, $\boldsymbol{\epsilon}_i$. For notational convenience we will refer to $\boldsymbol{\Sigma}_\mu$ and $\boldsymbol{\Sigma}_\epsilon$, rather than $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_B$. Typically zero-mean factor level priors are assumed; that is, $\boldsymbol{\beta}_0^{(b)} = \mathbf{0}$. Setting $\boldsymbol{\Psi}_b = \mathbf{0}$ and $\kappa_b = 0$ yields Jeffreys' noninformative prior $p(\mathbf{U}_b) \propto |\mathbf{U}_b|^{-(d+1)/2}$.

Although other covariance parameterizations may be chosen, $\mathbf{U}_b = \boldsymbol{\Sigma}_b + \frac{n_b}{n}\boldsymbol{\Sigma}_\epsilon$ is commonly used and should not be seen as arbitrary. In a scalar context Box and Tiao (1992, p. 252) have used $u_b = \sigma_b^2 + \frac{n_b}{n}\sigma_\epsilon^2$ since it appears in the decomposition of the likelihood for the one-way ANOVA model. The parameterization has also been utilized in the context of multivariate random effects by Everson and Morris (2000), since it is the conjugate prior. That is to say, the corresponding posterior will be of the same form. Further note that $\mathbf{U}_b$ is analogous to V_b of Section 2.1 with $I(b) = \{B\}$. Naturally, a prior may be specified directly on $\boldsymbol{\Sigma}_b$, although additional computation will then be needed.

3.2 Posterior Distributions

The distributions of (7)–(10), without additional specification, are not conducive to posterior sampling. In an MCMC setting this may manifest itself by failure to converge due to drifting in the parameter space (Gelfand and Sahu, 1999). This is analogous to non-identificability in the least squares setting, in which estimation of the set of all levels in a factor, $\{\boldsymbol{\beta}_1^{(b)}, \ldots, \boldsymbol{\beta}_{n_b}^{(b)}\}$, requires further constraints. By including analogous constraints in the prior distribution, a closed form of the posterior is possible, as well as for factorization between factors themselves. Degrees of freedom for each factor are then propagated

to the parameters of the corresponding covariance posterior. This parametrization is also beneficial in terms of computation since sets of factor levels are conditionally independent of one another. Using a vectorized form of the model, $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_n^T)^T \in \mathbb{R}^{nd}$ and $\boldsymbol{\beta}^{(b)} = (\boldsymbol{\beta}_1^{(b)^T}, \ldots, \boldsymbol{\beta}_{n_b}^{(b)^T})^T \in \mathbb{R}^{n_b d}$, is convenient for the development. The constraint $\mathbf{C}_b\boldsymbol{\beta}^{(b)} = \mathbf{0} \in \mathbb{R}^{c_b d}$, where there are c_b constraints, combined with (7), (8), is now

$$\mathbf{Y} \mid \{\boldsymbol{\beta}^{(b)}, \mathbf{C}_b\boldsymbol{\beta}^{(b)} = \mathbf{0}\}_{b=0}^{B-1}, \boldsymbol{\Sigma}_\epsilon \sim \mathrm{N}_{nd}\left(\sum_{b=0}^{B-1}\boldsymbol{\beta}^{(b)},\ \mathbf{I}_n \otimes \boldsymbol{\Sigma}_\epsilon\right), \tag{11}$$

$$\boldsymbol{\beta}^{(b)} \mid \mathbf{C}_b\boldsymbol{\beta}^{(b)} = \mathbf{0}, \boldsymbol{\Sigma}_b \sim \mathrm{N}_{n_b d}(\mathbf{1}_{n_b} \otimes \boldsymbol{\beta}_0^{(b)},\ \widetilde{\boldsymbol{\Omega}}_b), \qquad b = 0, \ldots, B, \tag{12}$$

with $\otimes$ denoting the Kronecker product. The rank-deficient $\widetilde{\boldsymbol{\Omega}}_b$, due to the constraint, causes (12) to be improper (see Appendix A.1 for the derivation of the density). The posterior, however, is proper. More over, when orthogonal factors are used, together with the constrained prior distributions, the full posterior from these densities can be efficiently factored into

$$p(\boldsymbol{\Sigma}_\epsilon,\ \{\boldsymbol{\Sigma}_b,\ \boldsymbol{\beta}^{(b)}\}_{b=0}^{B-1} \mid \mathbf{Y}) = p(\boldsymbol{\Sigma}_\epsilon \mid \mathbf{Y}) \prod_{b=0}^{B-1} p(\boldsymbol{\Sigma}_b \mid \mathbf{Y},\ \boldsymbol{\Sigma}_\epsilon)\ p(\boldsymbol{\beta}^{(b)} \mid \mathbf{Y},\ \boldsymbol{\Sigma}_\epsilon,\ \boldsymbol{\Sigma}_b), \tag{13}$$

where conditional independence of posteriors among the factors permits the partitioning. Details are given in Appendix A.2. Factored as such, there is no need for computationally intensive MCMC procedures. The corresponding distributions of (13) are

$$\boldsymbol{\Sigma}_\epsilon \mid \mathbf{Y} \sim W^{-1}\left(\boldsymbol{\Psi}_\epsilon + \sum_{i=1}^{n}(\mathbf{Y}_i - \widehat{\mathbf{Y}}_i)(\mathbf{Y}_i - \widehat{\mathbf{Y}}_i)^T,\ \kappa_\epsilon + n - \sum_{b=0}^{B-1} n_b\right), \tag{14}$$

$$\boldsymbol{\Sigma}_b + \frac{n_b}{n}\boldsymbol{\Sigma}_\epsilon \mid \mathbf{Y}, \boldsymbol{\Sigma}_\epsilon \sim W^{-1}\left(\boldsymbol{\Psi}_b + \sum_{j=c_b+1}^{n_b}(\widehat{\boldsymbol{\beta}}_j^{(b)} - \boldsymbol{\beta}_0^{(b)})(\widehat{\boldsymbol{\beta}}_j^{(b)} - \boldsymbol{\beta}_0^{(b)})^T,\ \kappa_b + n_b - c_b\right), \tag{15}$$

$$\boldsymbol{\beta}_j^{(b)} \mid \mathbf{Y}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\Sigma}_b \sim \begin{cases} \mathrm{N}_d\left(\widehat{\boldsymbol{\beta}}_j^{(b)},\ \frac{n_b}{n}\boldsymbol{\Sigma}_\epsilon\right) & j = 1, \ldots, c_b, \\ \mathrm{N}_d\left(\mathbf{P}_b^{-1}\boldsymbol{m}_j^{(b)},\ \mathbf{P}_b^{-1}\right) & j = c_b + 1, \ldots, n_b, \end{cases} \tag{16}$$

for $b = 0, \ldots, B-1$, where $\mathbf{P}_b = \boldsymbol{\Sigma}_b^{-1} + \frac{n}{n_b}\boldsymbol{\Sigma}_\epsilon^{-1}$, $\boldsymbol{m}_j^{(b)} = \frac{n}{n_b}\boldsymbol{\Sigma}_\epsilon^{-1}\widehat{\boldsymbol{\beta}}_j^{(b)} + \boldsymbol{\Sigma}_b^{-1}\boldsymbol{\beta}_0^{(b)}$ and where $\widehat{\cdot}$ denotes ordinary least-squares estimates. To sample from the full posterior, a realization from (14) is drawn, then realizations from (15) for each factor, and finally realizations of levels from (16) for each factor. Factor levels are adjusted according to constraints and sample covariances are calculated for each factor to obtain a realization of each

finite-population covariance, herein denoted by $\mathbf{S}_b$. The procedure provides a method of posterior sampling that is more formal than the ad-hoc approach reviewed in Section 2.2 while avoiding MCMC sampling.

3.3 Covariance Posterior Remarks

In all cases thus far covariances $\mathbf{\Sigma}_b$ are assumed to be of full rank. Hence, improper covariance posteriors will be due only to an insufficient number of observed factor levels. Díaz-García *et al.* (1997) offer a comprehensive look at all possible cases of improper Wishart distributions and following their terminology (15) would be classified as a pseudo-inverse-Wishart. Uhlig (1994) as well as Srivastava (2003) consider sampling with a pseudo-singular-Wishart distribution. Extending their work to inverse-Wishart distributions is one method for dealing with moderate discrepancies in the number of observed factor levels, $n_b < d$. Addressing cases in which $n_b \ll d$ is discussed in Section 6. Even in the remaining case, $n_b \geq d$, simulation from a posterior is not always efficient. Because support of the inverse-Wishart posterior requires positive definiteness in two respects, $\mathbf{\Sigma}_\beta > 0$, and $\mathbf{\Sigma}_\beta - \frac{n_b}{n}\mathbf{\Sigma}_\epsilon > 0$, the usual method of rejection sampling from an inverse-Wishart distribution is not always practical. Everson and Morris (2000) describe a more computationally efficient method to maintain positive definiteness through a Cholesky decomposition and maintaining positive eigenvalues while the sample realization is generated. In practice, as in the univariate case (Gelman, 2005), this is an issue when a variance parameter is hard to distinguish from zero. Problems will also occur when the dimension increases too quickly relative to sample sizes, an issue that is discussed in Section 6.

3.4 Multivariate Sources of Variation

Typical multivariate analysis of variance strategies rely on the distribution of the determinant of sums of squares matrices, e.g. Wilks' lambda distribution. However, multivariate sources of variability do not always lead to a single, clear criterion that indicates the greatest contributor to overall variability. For scalar variance components, $\sigma_b^2 > \sigma_\epsilon^2$ is clearly interpreted, however, due to the partial ordering of positive definite matrices, the analogous statement on covariance matrices is not useful. In other words, there is not a single, obvious comparison that can be made to determine which of two covariances are

"greater". Depending on the setting, there may exist an adequate scalar that sufficiently summarizes covariance characteristics. For the volume of ellipsoidal contours the determinant, $|\mathbf{\Sigma}|$, achieves this, while in other cases the sum of all entries, $\mathbf{1}^T\mathbf{\Sigma}\mathbf{1}$, or the sum of the marginal variances, $\text{tr}(\mathbf{\Sigma})$, may be appropriate. Scalar criteria with corresponding uncertainty intervals then allow multivariate sources of variability to be directly compared. Mardia *et al.* (1979) employed the determinant and trace, which correspond to the product and sum of eigenvalues; referring to them as the generalized variance and total variance, respectively. We amend the terminology, since inclusion of the sum of all matrix elements requires further delineation, by denoting $\mathbf{1}^T\mathbf{\Sigma}\mathbf{1}$ as the total variance, and $\text{tr}(\mathbf{\Sigma})$ as the total marginal variance.

Effectively relaying results of analysis of variance is one of the motivating factors that Gelman (2005) cites. Because the classical table of p-values does not yield any assessment on the magnitude of a factor's variance, the uncertainty of this assessment, nor a relative comparison between all factor variances, uncertainty intervals of all factor variances are the default choice for presenting results. Further, visual plots are convenient since they facilitate communication across disciplines. For explicit comparison of factors b and b', statements of the form $P\big(g(\mathbf{\Sigma}_b) > g(\mathbf{\Sigma}_{b'}) \mid \mathbf{Y}\big)$, utilizing an arbitrary matrix criterion $g(\cdot)$, may also be found.

4 Simulation Example

A toy example is now given in order to provide further background on the hierarchical Bayesian ANOVA method presented in the paper. Assessments of the uncertainty across all factors present in the statistical model are made, which immediately allow for relative comparisons. This is typically not possible with conventional ANOVA methods. Simulated data is used so that the conclusions can be evaluated by comparing the results to the true underlying parameters. Possible quantitative criteria of multivariate variability are also proposed in order to facilitate comparison of uncertainty from data in dimensions greater than one. While the variance parameters are assessed using their posterior distributions, we also consider approaches to summarize these multivariate distributions graphically. The visual summaries of this relative comparison are distinct from those made by Hawkins and Sutton (2009), and references therein, but convey the same idea.

The assumed model $\mathbf{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij} \in \mathbb{R}^d$, i.e. 3 factors with $b = 0, 1, 2$, in which observations are three-dimensional ($d = 3$). Parameters $\boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\epsilon$, and $\boldsymbol{\mu}$ are fixed. An individual observation is realized by generating $\boldsymbol{\alpha}_i$ and $\boldsymbol{\epsilon}_{ij}$, for $i = 1, \ldots, n_\alpha$, $j = 1, \ldots, n_\epsilon$, from mean-zero multivariate Gaussian distributions with their respective fixed covariances. The mean term $\boldsymbol{\mu}$ is added to the generated observation, yielding $n = n_\alpha n_\epsilon$ total observations.

Using (14)–(16), posterior distributions for each covariance criterion are obtained, and for three distinct simulation scenarios. Roughly speaking, the three scenarios can be explained by stating that variability introduced by factor $\boldsymbol{\alpha}$ is greater than (case 1), less than (case 2), or comparable to (case 3) variability introduced by errors $\boldsymbol{\epsilon}$. More specifically, the covariance matrix is decomposed into a vector of marginal standard deviations $\boldsymbol{s}_\alpha$ and a correlation matrix $\mathbf{R}_\alpha$, i.e. $\boldsymbol{\Sigma}_\alpha = \text{diag}(\boldsymbol{s}_\alpha)\mathbf{R}_\alpha\text{diag}(\boldsymbol{s}_\alpha)$. Throughout all simulations correlation matrices are fixed and $\boldsymbol{s}_\alpha$ is the only distinction between cases. Factor levels $\boldsymbol{\alpha}_i$ have the correlation structure $(\mathbf{R}_\alpha)_{1,2} = 0.3, (\mathbf{R}_\alpha)_{1,3} = 0.1, (\mathbf{R}_\alpha)_{2,3} = 0.5$. Errors $\boldsymbol{\epsilon}_{ij}$ have the covariance structure $(\boldsymbol{\Sigma}_\epsilon)_{i,j} = \rho^{|i-j|}$, and $\rho = 0.2$.

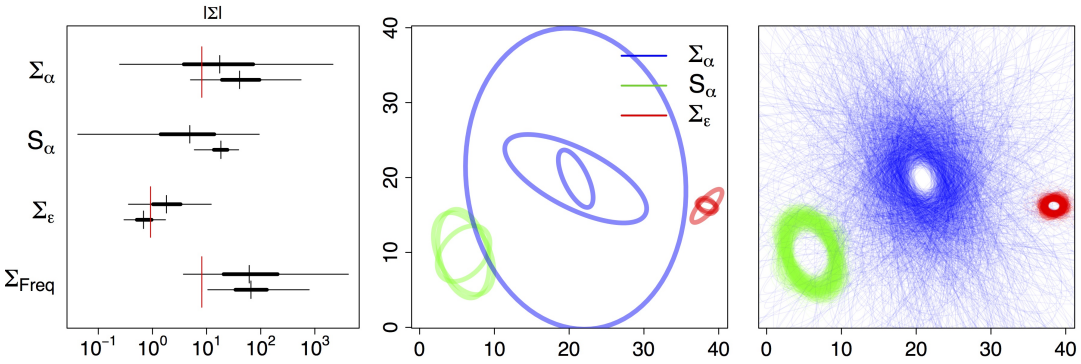

Figure 1: Simulation for two sample size pairs, $n_\alpha = 5, n_\epsilon = 3$, and $n_\alpha = 8, n_\epsilon = 5$. In both, factor $\boldsymbol{\alpha}$ varies greater than $\boldsymbol{\epsilon}$, with $\boldsymbol{s}_\alpha = (\sqrt{2}, \sqrt{2}, \sqrt{3})^T$. For uncertainty intervals, thin narrow lines denote endpoints at 95%, thicker middle portions denote 50%, and vertical ticks denote medians. Larger red vertical ticks denote true parameter values. Higher-positioned, wider intervals correspond to smaller factor level sample sizes, lower-positioned, narrower intervals correspond to larger factor level sample sizes (left). For second set of sample size pairs, ellipses from first two principal components are displayed for 2.5%, 50%, and 97.5% determinant-ordered percentiles of posterior distributions when 1000 posterior realizations have been drawn (center). All 1000 posterior ellipses (right).

The objective of the analysis is to assess the relative variability introduced by factor $\boldsymbol{\alpha}$ and by the factor $\boldsymbol{\epsilon}$, as well as the uncertainty in the assessment. Further, this is to be done in an appropriate multivariate context. Figure 1 displays the results of two simulation runs under the first scenario (case 1) using the determinant. In one simulation, the number of factor level realizations are $n_\alpha = 5, n_\epsilon = 3$ and in the second $n_\alpha = 8, n_\epsilon = 5$, which is to say less vs. more data. The left-most graph of Figure 1 displays uncertainty intervals for the determinant of super-population, finite-population, and error covariances. For factor $\boldsymbol{\alpha}$ the super-population and finite-population covariances are denoted as $\boldsymbol{\Sigma}_\alpha$ and $\mathbf{S}_\alpha$, respectively. To compare, a frequentist approach to obtain a confidence interval has also been used, $\boldsymbol{\Sigma}_{\text{Freq}}$, and is described in Appendix B. Narrow lines denote 95% quantiles, thicker lines 50%, and a vertical tick is placed at the median. The upper set of intervals, which are intuitively wider, correspond to less data, while the lower set of intervals correspond to more. By vertical comparison of the uncertainty intervals, we see that for $n_\alpha = 5, n_\epsilon = 3$ all intervals overlap, and hence no distinction can be made between the sources of variability. For $n_\alpha = 8, n_\epsilon = 5$ however, there is no overlap of the uncertainty intervals, suggesting that both super-population variability, and the finite-population variability are greater than error variability.

Although visual representations of multivariate variability are needed in order to communicate results to a broad audience, there is still no widely accepted way to do so. Tokuda *et al.* (2011), and references therein, provide some options to visualizing covariance matrices. The remaining two plots of Figure 1 go slightly further by presenting a diagnostic look at the uncertainty in the covariance assessment. The center graph displays ellipses from first two principal components corresponding to $2.5\%, 50\%$, and 97.5% determinant-ordered percentiles of the posterior distributions, specific values of which can be seen on endpoints of the uncertainty intervals. The right-most graph goes one step further by showing the uncertainty of the assessment of the variability. This is done by plotting of all 1000 ellipses from each posterior distribution to see the size, shape, and orientation, of all factor covariances. Size renders an idea of the magnitude of the marginal variances. The disparity in size between the ellipses corresponding to $\mathbf{S}_\alpha$ and $\boldsymbol{\Sigma}_\epsilon$ suggest that the marginal variances of the former are greater than those of the latter. Through shape, some insight into the dependence of a factor's covariance may be gleaned. Lastly, the orientation, or varying orientation, suggests the uncertainty of the

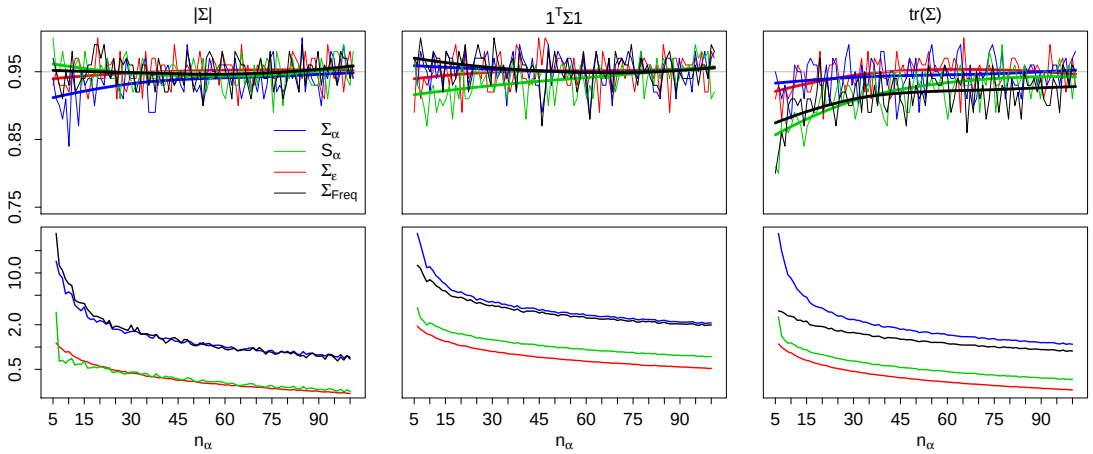

Figure 2: Coverage is estimated with 100 simulations for values of $n_\alpha = 5, \ldots, 100$ and for which the sources of variability are approximately equal (case 3). Coverage estimates are given for $\mathbf{\Sigma}_\alpha$, $\mathbf{S}_\alpha$, $\mathbf{\Sigma}_\epsilon$, as well as for a frequentist based confidence interval, $\mathbf{\Sigma}_{\text{Freq}}$. Gray lines denote 95% nominal coverage while bold lines show LOESS smoothed coverages (top). Average uncertainty interval widths are shown on a log scale (bottom).

dependence, e.g. as the orientation of the ellipses corresponding to $\mathbf{\Sigma}_\alpha$ fluctuate greatly, there is not much that can be said about its dependence structure.

To gain insight into the coverage success and uncertainty interval widths, we have carried out 100 simulations over different values of n_α and each of the different scenarios of variability sources. For each of these simulations the number of replicates is fixed with $n_\epsilon = 15$. Figure 2 displays results for the scenario in which variability sources are comparable (case 3). Results in all three scenarios were relatively similar with respect to coverage, noting that uncertainty interval widths increase as the magnitude of variability increases. Coverage and interval widths for $\mathbf{\Sigma}_\alpha$ and $\mathbf{\Sigma}_{\text{Freq}}$ should be compared as they correspond to the same true, unknown covariance.

5 Application

The method defined in Section 3 is now applied to a bivariate dataset of global temperature (°C) and precipitation (mm/day) projections for 9 decadal averages of boreal summer months, June, July, August, during the remaining century. The data is part of the CMIP3 project (Meehl *et al.*, 2000) and has been used in the framework of the Fourth Assessment Report (AR4) for the Intergovernmental Panel on Climate Change (IPCC) as well. The

first factor in the model consists of 13 levels, each representing a single atmosphere-ocean general circulation model (GCM). Specific GCMs represented in the data are CNRM-CM3, CSIRO-MK3.0, GFDL-CM2.0, GFDL-CM2.1, GISS-MODEL-ER, INM-CM3.0, IPSL-CM4, MIROC3.2, ECHAM5/MPI-OM, MRI-CGCM2.3.2A, CCSM3.0, PCM, and UKMO-HadCM3 (see Meehl *et al.*, 2007 for detailed descriptions). Although some of these have been simulated for several runs, only one run from each has been used. The second factor covers 3 greenhouse gas emissions scenarios (SCN) that have been defined by the Special Report on Emissions Scenarios, and are identified as A1B, A2, and B1 (Nakićenović and Swart, 2000). The raw data, partitioned by SCN, is shown in Figure 3. There are additional GCMs have been developed, however, only those for which complete data was available for each SCN were used.

One fundamental objective of the analysis is then to compare how these factors contribute to overall variability of global climate averages, how they relate to one another, and what the uncertainty in this assessment is. Further, we will consider the uncertainty of relative change between initial and final decadal intervals. This will be done from both a finite and super-population point of view.

Bias and dependence among climate models is an issue that has in recent years begun to be examined further, beginning with Tebaldi and Knutti (2007), Jun *et al.* (2008),

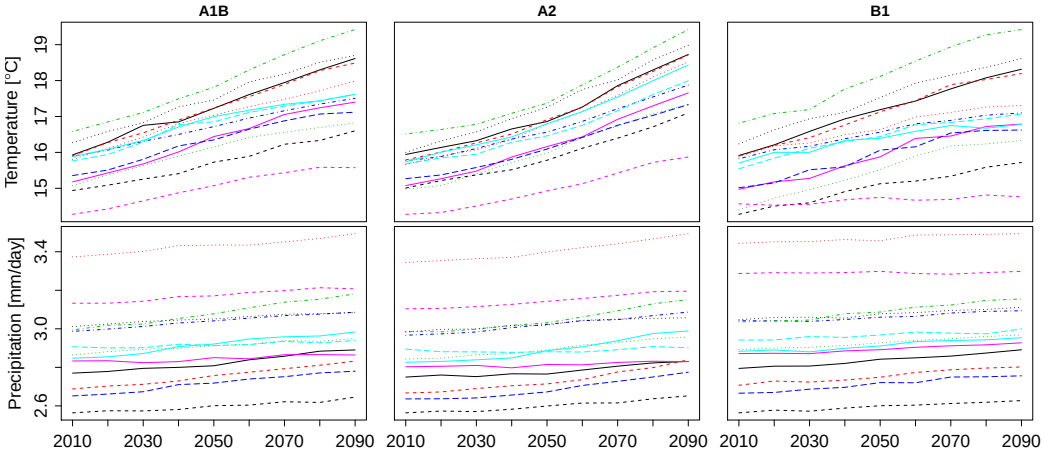

Figure 3: Global averages for temperature (top) and precipitation (bottom) over nine decadal intervals ($n_t = 9$), with thirteen climate models ($n_\alpha = 13$) represented in different colors and line types, and with three emissions scenarios ($n_\beta = 3$) in columns. Because of similarities across columns, the most distinguishing factor is suggested to be GCM.

Knutti *et al.* (2010), and references therein. While we adopt the statistical assumption that has often been used when working with sets of GCMs, which is to assume that they are independently drawn from a large or infinite population of climate models, it is possible to incorporate dependencies among the GCMs into their prior joint distribution. We foresee this being done by specifying the covariance $\mathbf{\Omega}_b$ of (12) with quantified dependencies between GCM models, such as those found by Jun *et al.* (2008), rather than the independence implied by $\mathbf{I}_{n_b}$ in (12). However, by using the prior independence assumption, it is straightforward to consider the uncertainty of a novel GCM. Following in this manner, uncertainty for each source of variation in the model is considered, and uncertainty in an unobserved, novel factor level is also evaluated. Under these assumptions the method may be regarded, at the very least, as an exploratory tool. Preliminary analysis steps have suggested the model

$$\mathbf{Y}_{ijt} = \boldsymbol{\mu}_0 + \boldsymbol{\alpha}_{0,i} + \boldsymbol{\beta}_{0,j} + \boldsymbol{\gamma}_{ij} + x_{1,t}\boldsymbol{\mu}_1 + x_{1,t}\boldsymbol{\alpha}_{1,i} + x_{1,t}\boldsymbol{\beta}_{1,j} + x_{2,t}\boldsymbol{\mu}_2 + \boldsymbol{\epsilon}_{ijt}, \tag{17}$$

where $i = 1, \ldots, n_\alpha$, $j = 1, \ldots, n_\beta$, $t = 1, \ldots, n_t$, with $n_\alpha = 13, n_\beta = 3, n_t = 9, n = n_\alpha n_\beta n_t$, and $d = 2$. Time covariate x_1 is centered such that $x_{1,t} = -4, \ldots, 4$. A quadratic time covariate is also used, x_2, and is transformed to be orthogonal to other predictors in the model. Factors of interest are GCM, $\boldsymbol{\alpha}$, and SCN, $\boldsymbol{\beta}$, and their interaction, $\boldsymbol{\gamma}$. The first two are further delineated as a constant (varying intercepts), $\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0$, and temporally (varying slopes), $\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1$.

Posterior distributions of factors $\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0$, and $\boldsymbol{\gamma}$ are derived from (15) and (16). Factors $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ differ slightly as they correspond to the regression model formulation. A multivariate factor associated with a covariate would, in general, be multiplied by a matrix, e.g. $\mathbf{X}_{1,t}$. Utilizing (23), and accounting for implicit constraints placed on the parameters, super-population and factor level posteriors of factor $\boldsymbol{\alpha}_1$ are then

$$\boldsymbol{\Sigma}_{\alpha_1} + \mathbf{V}_{\alpha_1} \mid \mathbf{Y}, \boldsymbol{\Sigma}_\epsilon \sim W^{-1}\left(\sum_{i=c_{\alpha_1}+1}^{n_\alpha} \widehat{\boldsymbol{\alpha}}_{1,i}\widehat{\boldsymbol{\alpha}}_{1,i}^T,\ n_\alpha - c_{\alpha_1}\right), \tag{18}$$

$$\boldsymbol{\alpha}_{1,i} \mid \mathbf{Y}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\Sigma}_{\alpha_1} \sim \begin{cases} \mathrm{N}_d\left(\widehat{\boldsymbol{\alpha}}_{1,i},\ \frac{n_\alpha}{n}\boldsymbol{\Sigma}_\epsilon\right) & i = 1, \ldots, c_{\alpha_1}, \\ \mathrm{N}_d\left(\mathbf{P}_{\alpha_1}^{-1}\boldsymbol{m}_{\alpha_1,i},\ \mathbf{P}_{\alpha_1}^{-1}\right) & i = c_{\alpha_1}+1, \ldots, n_\alpha, \end{cases} \tag{19}$$

where $\mathbf{V}_{\alpha_1} = \frac{1}{n_\beta}(\sum_{t=1}^{n_t} \mathbf{X}_{1,t}^T\boldsymbol{\Sigma}_\epsilon^{-1}\mathbf{X}_{1,t})^{-1}$, $\mathbf{P}_{\alpha_1} = \boldsymbol{\Sigma}_{\alpha_1}^{-1} + \mathbf{V}_{\alpha_1}^{-1}$, and $\boldsymbol{m}_{\alpha_1,i} = \mathbf{V}_{\alpha_1}^{-1}\widehat{\boldsymbol{\alpha}}_{1,i}$. For model (17) the covariate matrix is $\mathbf{X}_{1,t} = \mathrm{diag}(x_{1,t})$, and thus $\mathbf{V}_{\alpha_1} = (n_\beta \sum_{t=1}^{n_t} x_{1,t}^2)^{-1}\boldsymbol{\Sigma}_\epsilon$. The posterior of factor $\boldsymbol{\beta}_1$ is found similarly.

Figure 3 suggests that GCM is the most distinguishing feature. Figure 4 confirms this assessment as $\boldsymbol{\alpha}_0$ is seen to be the most significant source of variability among all factors for both super- and finite-population cases. Further conclusions that might be foreseen using Figure 3 are that $\boldsymbol{\alpha}_1$ contributes little variation, since slopes across GCMs are fairly consistent; and that interaction $\boldsymbol{\gamma}$ is minor, because ranking of GCMs from one SCN to the next is fairly consistent. This is confirmed by comparing the estimated variance components for $\boldsymbol{\alpha}_1, \boldsymbol{\gamma}$ with that of $\boldsymbol{\epsilon}$ in Figure 4. Comparison of these two figures with one another illustrates the additional uncertainty of super-population parameters over their finite-population counterparts. Reiterating, uncertainty intervals for the super-population covariance criteria are wide because they account for uncertainty of all factor levels, even those that are not part of the observed data. The uncertainty is particularly large when there are few observed factor levels, as in the case of SCN factors $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1$. Finite-population covariance uncertainty intervals are generally smaller, because they are concerned with variability of only the factor levels that have been observed. Note that while the data uses familiar units of temperature and precipitation, the scale of Figure 4 is not easily interpretable, e.g. covariance posterior determinants are in units of $(^\circ\text{C})^2 \cdot (\text{mm/day})^2$. The analysis has been also carried out using the standardized data,

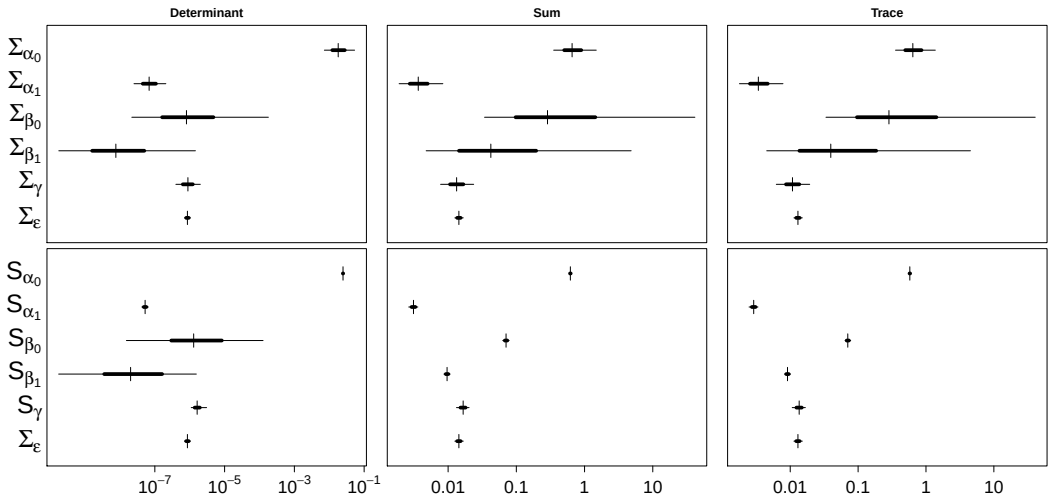

Figure 4: Uncertainty intervals for summaries of super-population variances $\boldsymbol{\Sigma}_b$ (top) and finite-population variances $\mathbf{S}_b$ (bottom) on a log scale using generalized variance (left), total variance (middle), and total marginal variance (right). Nominal coverages of 95% (thin lines) and 50% (thick lines), and the median (vertical line) are denoted using quantiles of the corresponding posterior distributions.

such that observed temperature and precipitation each have sample mean and standard deviation of 0 and 1, respectively. The relative comparisons of the factor variances under the standardized units was, however, nearly indistinguishable. Since focus is on relative comparison across sources of variability, the more familiar units have been kept.

Posterior predictive distributions, often used to perform model checking and diagnostics, can also be utilized to identify distinct sources of variability. The posterior predictive is conditional on observations with levels from each factor assumed to be, a) the same as those factor levels that have been observed data, b) unobserved/novel factor levels. Figure 5 examines posterior density $p(\widetilde{\mathbf{Y}}_{ijn_t} - \widetilde{\mathbf{Y}}_{ij1}|\{\mathbf{Y}_{ijt}\})$, the difference between posterior predictive distributions at the final, $t = n_t$, and initial, $t = 1$, decades, i.e. $2090 - 2010$. The figure shows that uncertainty of a novel SCN level is greater than that of a novel GCM level. This is extracted by focusing on temporal factors, $\boldsymbol{\alpha}_{1i'}$, $\boldsymbol{\beta}_{1j'}$, where indices i', j' signify new, unobserved factor levels. Linear and quadratic terms, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ are included, although additional (but negligible) variability from these terms has been disregarded. In the left-most plot, uncertainty from all three observed SCN levels is assumed. Therefore variability from each SCN level posterior is included, as described by a). A new unobserved GCM level is assumed, which utilizes the super-population posterior, as in b). The center plot of Figure 5 assumes three observed GCM factor levels

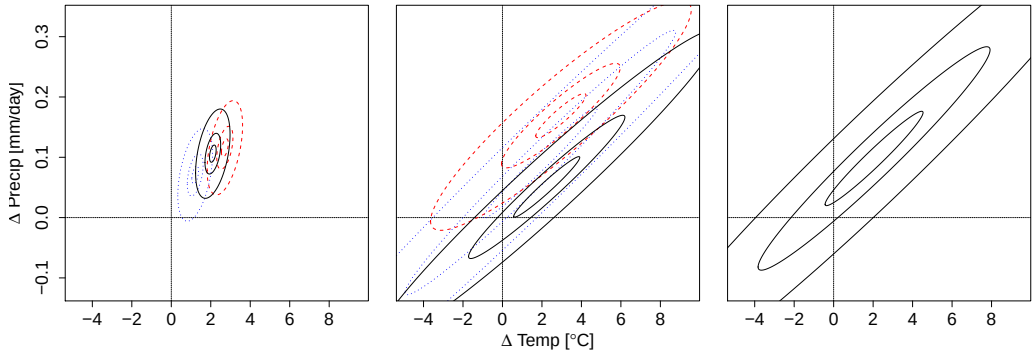

Figure 5: Posterior predictive distribution of relative change between 2090 and 2010 with unobserved GCM level and observed SCN levels A1B, A2, B1 (black solid, red dashed, blue dotted) (left); with observed GCM levels GFDL-CM2.0, ECHAM5/MPI-OM, PCM (black solid, red dashed, blue dotted) and unobserved SCN level (center); and with both GCM and SCN unobserved levels (right). Contours correspond to $5\%, 25\%$, and 75% percentiles.

(GFDL-CM2.0, ECHAM5/MPI-OM, and PCM) and have been selected so as to best represent the range of all the observed GCM levels. Additionally an unobserved SCN level is assumed. In this case, the high level of uncertainty in the SCN super-population variance is propagated through to the predictive distribution. The high level of uncertainty may be attributed more to the limited number of observed SCN levels, $n_\beta = 3$, than to unusually high variability between these factor levels themselves. The right-most panel, in which an unobserved level for both GCM and SCN is assumed, naturally shows the largest degree of uncertainty.

Figure 6 is similar to 5 except that temperature and precipitation changes are shown over all time differences, $2020-2010, \ldots, 2090-2010$, thereby highlighting the increasing uncertainty over time. As with the previous figure, an estimate of the internal variability is, by construction, included in the uncertainty for relative change. This should be regarded as only a rough estimate though since decadal, rather than annual, data has been modeled. Again, the three cases examined are for an unobserved GCM level and observed SCN level (A1B), observed GCM level (ECHAM5/MPI-OM) and unobserved SCN level,

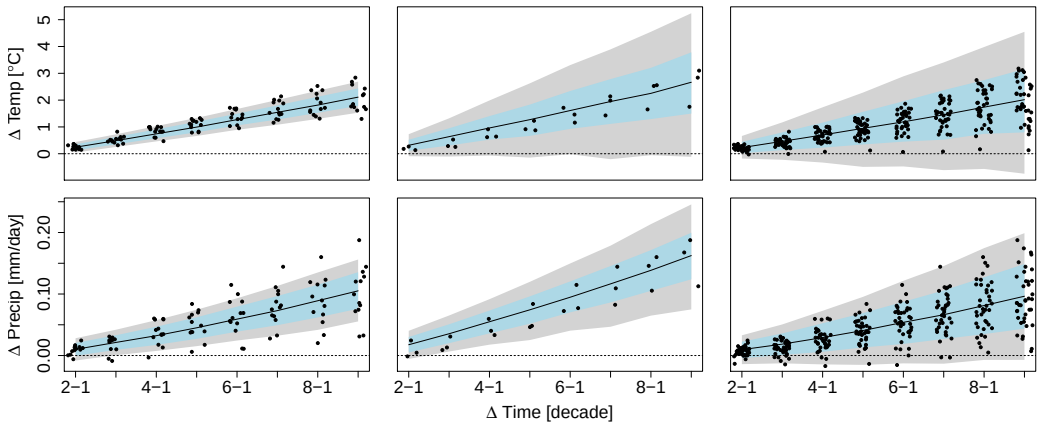

Figure 6: Posterior predictive distributions for uncertainty bounds of change in temperature and precipitation over decadal differences $2020-2010, \ldots, 2090-2010$ with unobserved GCM level and observed SCN level A1B (left); with observed GCM level ECHAM5/MPI-OM and unobserved SCN level (center); and with both GCM and SCN as unobserved levels (right). Grey areas correspond to regions of 75% uncertainty (i.e. 12.5% and 87.5% percentiles), blue areas to 50%, and solid lines to medians. Plotted points (jittered) show observational differences for corresponding factor levels.

and both unobserved GCM and unobserved SCN level, respectively from left to right. The left-most figure may be compared to the multi-model average plot of Meehl *et al.* (2007, Fig. SPM.5) but differs in that we assume a novel, unobserved GCM level, and thus additional uncertainty. As with the previous figure, the middle column of Figure 6 highlights the high uncertainty when a novel SCN level is considered.

6 Discussion

More and more, researchers are utilizing analysis of variance approaches that allow appropriate parameters, e.g. super- or finite-population, to be used. In this paper, technical details of a hierarchical Bayesian approach towards ANOVA have been derived, thereby allowing the method to be used in the analysis of simulated climate projections. More generally, this has been extended to multivariate cases and may also yield inferential procedures on arbitrary covariance criteria. Further extensions to the methodology must explicitly address increasing dimensionality. For moderately sized dimensions d, relative to number of observations, improper inverse-Wishart distributions, and/or priors that impose particular dependence structures, are possible options. For cases in which d is very large, stricter covariance assumptions may be employed. In the spatial context, properties such as stationarity allow covariance parameter space to be reduced, e.g. range, sill, and nugget in a spatial covariance function. Because simultaneous estimation of such parameters is nontrivial, some parameters are often assumed, or estimated empirically in earlier analysis steps, as in Furrer *et al.* (2007). In other cases, so as to maintain computational feasibility, sparsity restrictions are placed on covariances (Cressie and Johannesson, 2008; Furrer *et al.*, 2006; Stein, 2008). For many such scenarios a covariance is decomposed into a correlation matrix and a scalar variance parameter. The method is then carried out with the inverse-Wishart posterior density transformed through a spectral decomposition of the correlation matrix, thus allowing for efficient posterior sampling for cases in which $d \gg n$. This extension, as carried out by Furrer *et al.* (2012), offers an alternative to geostatistical model analyses that have previously relied on computationally intensive MCMC methods. Other difficulties encountered are unbalanced designs and linearly dependent predictors. MCMC, or approximation methods such as INLA (Rue *et al.*, 2009), are possible solutions to both of these problems, and are areas currently

under investigation.

Acknowledgments

We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modeling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy. SG and RF acknowledge the University Research Priority Program (URPP) at the University of Zurich. SS acknowledges the National Center for Atmospheric Research (NCAR), which is managed by the University Corporation for Atmospheric Research (UCAR) under the sponsorship of the U.S. National Science Foundation (NSF).

References

Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*. Wiley-Interscience.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc. B*, **70**, 209–226.

Díaz-García, J. A., Gutierrez Jáimez, R., and Mardia, K. V. (1997). Wishart and pseudo-Wishart distributions and some applications to shape theory. *J. Multivariate Anal.*, **63**, 73–87.

Everson, P. J. and Morris, C. N. (2000). Simulation from Wishart distributions with eigenvalue constraints. *J. Comp. Graph. Statist.*, **9**, 380–389.

Furrer, R., Geinitz, S., and Sain, S. R. (2012). Assessing variance components of general circulation model output fields. *Environmetrics*, **23**, 440–450.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.*, **15**, 502–523.

Furrer, R., Sain, S. R., Nychka, D. W., and Meehl, G. A. (2007). Multivariate Bayesian analysis of atmosphere-ocean general circulation models. *Environmental and Ecological Statistics*, **14**, 249–266.

536 Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, improper priors, and gibbs sampling
537 for generalized linear models. *J. Am. Stat. Ass.*, **94**, 247–253.

538 Gelman, A. (2005). Analysis of variance: Why it is more important than ever. *Ann.*
539 *Statist.*, **33**, 1–31.

540 Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multi-*
541 *level/Hierarchical Models*. Cambridge University Press, 1 edition.

542 Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their
543 applications. *Biometrika*, **57**, 97–109.

544 Hawkins, E. and Sutton, R. (2009). The potential to narrow uncertainty in regional
545 climate predictions. *Bull. Am. Met. Soc.*, **90**, 1095–1107.

546 Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables.
547 *Biometrika*, **48**, 419–426.

548 Jun, M., Knutti, R., and Nychka, D. W. (2008). Spatial analysis to quantify numerical
549 model bias and dependence: How many climate models are there? *J. Am. Statist.*
550 *Ass.*, **103**, 934–947.

551 Kaufman, C. G. and Sain, S. R. (2010). Bayesian functional ANOVA modeling using
552 Gaussian process prior distributions. *Bayes. Anal.*, **5**, 847–874.

553 Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A. (2010). Challenges in
554 combining projections from multiple climate models. *J. Clim.*, **23**, 2739–2758.

555 Kreft, I., De Leeuw, J., and de Leeuw, J. (1998). *Introducing multilevel modeling*. Sage
556 London.

557 Leinonen, T., O'Hara, R. B., Cano, J. M., and Merilä, J. (2008). Comparative studies
558 of quantitative trait and neutral marker divergence: a meta-analysis. *J. Evolutionary*
559 *Bio.*, **21**, 1–17.

560 Lencina, V. and Singer, J. (2006). Measure for measure: exact f tests and the mixed
561 models controversy. *Int. Statist. Review*, **74**, 391–402.

562 Lencina, V. B., Singer, J. M., and Stanek, E. J. (2005). Much ado about nothing: the
563 mixed models controversy revisited. *Int. Statist. Review*, **73**, 9–20.

564 Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs - a bayesian
565 modelling framework: Concepts, structure, and extensibility. *Stat. and Comp.*, **10**,
566 325–337.

567 Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic
568 Press.

569 Meehl, G. A., Boer, G. J., Covey, C., Latif, M., and Stouffer, R. J. (2000). The coupled
570 model intercomparison project (CMIP). *Amer. Met. Soc. Bulletin*, **81**, 313–318.

571 Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, P., Gaye, A. T., Gregory,
572 J. M., Kitoh, A., Knutti, R., Murphy, J. M., Noda, A., Raper, S. C. B., Watterson,
573 I. G., Weaver, A. J., and Zhao, Z.-C. (2007). Global climate projections. In Solomon,
574 S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and
575 Miller, H. L., editors, *Climate Change 2007: The Physical Science Basis. Contribution*
576 *of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel*
577 *on Climate Change*. Cambridge: Cambridge University Press.

578 Nakićenović, N. and Swart, R., editors (2000). *Special Report on Emission Scenarios*.
579 Intergovernmental Panel on Climate Change, Cambridge University Press.

580 Nelder, J.*et al.* (1998). The great mixed-model muddle is alive and flourishing, alas! *Food*
581 *quality and preference.*, **9**, 157.

582 Nelder, J. A. (1977). A reformulation of linear models. *J. R. Statist. Soc. A*, **140**, 48–77.

583 Nelder, J. A. (1994). The statistics of linear models: back to basics. *Statist. and Com-*
584 *puting*, **4**, 221–234.

585 Nelder, J. A. (2008). What is the mixed-models controversy? *Int. Statist. Review*, **76**,
586 134–135.

587 Qian, S. S. and Shen, Z. (2007). Ecological applications of multilevel analysis of variance.
588 *Ecology*, **88**, 2489–2495.

Paper III

Bayesian analysis of variance for relative assessment of general circulation model effect

Steven Geinitz & Reinhard Furrer & Stephan R. Sain

Paper under revision for resubmission to the
International Journal of Climatology.

Bayesian multilevel analysis of variance for relative comparison across sources of global climate model variability

Steven Geinitz[1], Reinhard Furrer[1], Stephan R. Sain[2]

[1]Institute of Mathematics, University of Zurich, Zurich, Switzerland

[2]Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, Boulder CO, USA

Projections of future climate conditions are carried out by many research institutions, each with their own general circulation model to do so. The projections are additionally subjected to distinct anthropogenic forcings, specified by future greenhouse gas emissions scenarios. These two factors, together with their temporal effects and interaction, create several potential sources of variation in final climate projection output. Multilevel statistical models, and specifically multilevel ANOVA, have come to be widely used for many reasons, not least of which is their ability to comprehensively assess many different sources of variation. In this paper a Bayesian multilevel ANOVA approach is applied to climate projections to assess each of these sources of variation, estimate the uncertainty regarding the assessment, and to allow comparison across all sources. The data originate from phase three of the Coupled Model Intercomparison Project, consisting of 15 circulation models and 3 emissions scenarios over 9 decadal time periods for boreal summer and winter. A single observation is a spatial field of 2.8 degree resolution with units of precipitation in mm/day, and temperature in degrees Celsius. As this approach towards ANOVA is relatively novel, and particularly so for spatial data, a short discussion of conventional ANOVA and the new methodology is provided.

Keywords: climate change; uncertainty; variance components

1 Introduction

An atmosphere-ocean general circulation model (GCM) attempts to simulate the Earth's climate system, including the complex interactions among the atmosphere, land surface, ocean, sea ice and other biogeochemical processes. In addition to the profound scientific, social and political impacts (Hansen *et al.*, 2006) that results from these GCMs carry on

the grand scale, they are also influencing high-resolution, regional climate modeling efforts (Kaufman and Sain, 2010; Jin *et al.*, 2010; Forland *et al.*, 2011). Of the approximately twenty models in use, they mostly differ in how small scale processes are implemented and in the numerical approaches used to solve the necessary systems of partial differential equations. Together, these twenty GCMs constitute one potential source of variation in climate projection output. Another potential source is the level and composition of greenhouse gas emissions that is predicted to occur. Emissions scenarios (ES) on the quantity and composition of these emissions have been defined (Nakićenović and Swart, 2000) and included in climate modeling runs, to assess their effect. Although non-trivial, ascertaining the variability that is introduced into climate projections from these two sources, and others, has the potential to offer a new perspective on climate modeling.

In this paper we investigate how these two factors, GCM and ES, contribute to variability in a set of climate projections over summer and winter periods for both precipitation and temperature. The data, which has been made available as part of the third phase of the Coupled Model Intercomparison Project (CMIP3) (Meehl *et al.*, 2007a), is analyzed using a relatively novel approach towards analysis of variance. The analysis yields an assessment for each source of variation, estimates the uncertainty regarding each assessment, and further allows for relative comparison across all sources.

In Section 2 we describe the analysis of variance methodology to be used and illustrate it with an elementary example. This section may be skipped by those familiar with analysis of variance, although the example provided can be seen as an abstract simplification of the CMIP3 data later examined. For more explicit details on the statistical methodology the reader is referred to Furrer *et al.* (2012), in which a subset of only the temperature data presented here was used. Section 3 describes the data in detail, defines the statistical model, and summarizes the results. Section 4 concludes with a discussion of extensions and generalizations to be addressed in the future.

2 Methodology

2.1 Background

Researchers often recognize statistics as a means of inference for estimation, prediction, decision making, etc. Although at its roots, statistics aims to recognize and quantify

sources of uncertainty. In other words, analyzing variability, or analysis of variance (ANOVA). Broadly speaking there are two points of view under the umbrella of ANOVA that can be taken to examine the effect of various factors, which differentiate observations from one another. The first is commonly employed when factors are considered fixed, that is when all factor outcomes are included in the observations. The second, variance component analysis, often implies that effects are random, and that one must consider effects of the factor that have not been observed. This description is, however, a broad generalization, as there is no clear consensus on the precise definition of fixed and random effects (Gelman, 2005). The researcher must then pigeonhole their scientific inquiries into the context of one of these two types of analyses. Gelman (2005) has proposed a more harmonious, comprehensive solution through the use of a novel formulation of analysis of variance within the Bayesian paradigm. We follow this formulation, which allows one to restate fixed and random effects into more intuitive finite and super population variances.

As a toy example consider a group of three lab technicians, each of which carries out an experiment nine times, resulting in a total of 27 observations, e.g. yield of a synthesized solution. There are two potential sources of variation, lab technician and experiment error. To answer the question of whether there is significant variability introduced by the lab technician, the scenario may be represented by model $Y_{ij} = \beta_i + \epsilon_{ij}$, with $i = 1, \ldots, 3$ and $j = 1, \ldots, 9$. Conventional ANOVA, which corresponds to a fixed effect approach, would address the question by utilizing a sums of squares decomposition. The scenario is simulated with β_i's and ϵ_{ij}'s being generated from mean zero Gaussian distributions with variances $\sigma^2_\beta = 5$ and $\sigma^2_\epsilon = 3$, respectively. Classical, frequentist ANOVA assumes that

	Df	Sum Sq	Mean Sq	F stat	p-value
Betas	3	105.824	35.275	14.571	1.3×10^{-5}
Residuals	24	58.099	2.421		

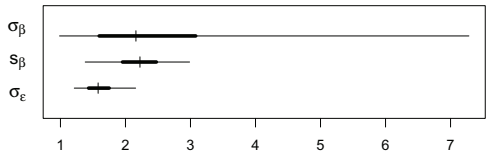

Figure 1: Classical sums of squares ANOVA summary culminates with p-value corresponding to hypothesis of equal beta effects (left). Bayesian interpretation of ANOVA displaying credible intervals for standard deviation components (right). Narrow lines denote posterior credible intervals for 95% (2.5% and 97.5% quantiles), thicker lines for 50% credible intervals (25% and 75% quantiles), and vertical mark is at the median (50% quantile). True simulation values are $\sigma_\beta = \sqrt{5} \approx 2.2$ and $\sigma_\epsilon = \sqrt{3} \approx 1.7$.

the lab technician effect is fixed, then tests the hypothesis that these effects are all equal, i.e. $H_0 : \beta_1 = \beta_2 = \beta_3$. In the simulation the hypothesis is rejected (at any reasonable level of significance, including 0.05) with a p-value of 0.000013, as seen in the left side of Figure 1. Thus, lab technician effect is deemed to be statistically significant. If lab technicians are assumed to come from a larger population, then a random effect variance component approach may be chosen and a separate analysis is done.

The Bayesian ANOVA strategy simultaneously addresses both of these questions. The super population variance, equivalent to the random effect variance component, pertains to the larger population. Because this includes any potential future lab technician, the uncertainty of inferential statements for the super population must include the uncertainty of unobserved lab technicians. Hence, the wider interval for σ_β in the right side of Figure 1. The finite population variance does not correspond directly to any term in conventional ANOVA, rather it is the variance s_β^2 of these three particular lab technicians. Since the uncertainty of future lab technicians is excluded, inferential statements on the finite population variance will be more "precise", hence the narrower interval for s_β in Figure 1.

Statements, roughly analogous to conventional ANOVA hypotheses, in the form of probabilities on population variances, or standard deviations, can also be made. From the example, the probability that variability between these three particular lab technicians is greater than a value c, where $c = 0.5$ is assumed to be scientifically relevant, e.g. a threshold for laboratory regulations, is $P[s_\beta > c] = 0.9997$, conditional on the observations. Relative comparison, not possible with conventional ANOVA, offers greater perspective in assessing the sources of variability. Namely, the probability that variability from these three particular lab technicians is greater than variability from measurement errors is $P[s_\beta > \sigma_\epsilon] = 0.8906$, conditional on the observations. Summarizing, the method utilizes the uncertainty prescribed by the posterior distributions to assess "significance" of the variances in both an absolute and relative sense.

In general, one must consider any number of factors. From the illustration provided these could be, for example, different machines used in the experiment, or time of day that the experiment is conducted, etc. To facilitate this it is useful to refer to each factor as a *batch*, that is as a set of varying coefficients, or effects. In this way, overall mean, factors, their interactions, terms with covariates, as well as error terms can be referred

to similarly. A Bayesian multilevel framework is then used to model each batch of effects together with its variance parameter. For the previous example with an overall mean μ data observations are modeled as $Y_{ij} = \mu + \beta_i + \epsilon_{ij}$, and a Bayesian hierarchy is assigned to each of the three batches

$$\mu \mid \sigma_\mu^2 \sim N(\mu_0, \sigma_\mu^2), \qquad \beta_i \mid \sigma_\beta^2 \overset{iid}{\sim} N(\beta_0, \sigma_\beta^2), \qquad \epsilon_{ij} \mid \sigma_\epsilon^2 \overset{iid}{\sim} N(\epsilon_0, \sigma_\epsilon^2),$$
$$\sigma_\mu^2 \sim F_\mu, \qquad \sigma_\beta^2 \sim F_\beta, \qquad \sigma_\epsilon^2 \sim F_\epsilon,$$

for $i = 1, \ldots, n_I,\ n_j = 1, \ldots, n_J$. Batch effects are each assigned a Gaussian distribution $N(\mu, \sigma^2)$, with mean and variance denoted by μ and σ^2, respectively. In this case, a general probability distribution F has been assigned to the batch variance components, although the inverse-gamma distribution is a common choice for its conjugate properties. For a wide class of linear models, such as the previous example, posterior densities of super population variances, batch effects, and finite population variances are derived. Specifically, the full joint posterior is factored to obtain a conditional posterior on each parameter of interest. Further, when orthogonal designs are available conditional posteriors are independent so that the need for MCMC methods is obviated. In the example above, the posterior of super population variances and batch effects is

$$\mathrm{p}(\sigma_\epsilon^2, \sigma_\beta^2, \{\beta_i\} \mid \{Y_{ij}\}) = \mathrm{p}(\sigma_\epsilon^2 \mid \{Y_{ij}\}) \cdot \mathrm{p}(\sigma_\beta^2 \mid \{Y_{ij}\}, \sigma_\epsilon^2) \cdot \mathrm{p}(\{\beta_i\} \mid \{Y_{ij}\}, \sigma_\epsilon^2, \sigma_\beta^2), \qquad (1)$$

where $\mathrm{p}(\theta \mid \{Y_i\})$ denotes a conditional probability density,"parameter θ given the set of observations $Y_1, \ldots, Y_n$". Posterior distributions of finite population variances, which are a function of the batch effects e.g. $s_\beta^2 = s_\beta^2(\{\beta_i\})$, are then calculated (Gelman, 2005).

Prediction may be carried out with the posterior predictive density $\mathrm{p}(\widetilde{Y} \mid \{Y_{ij}\})$, where $\widetilde{Y}$ represents a future observation. One may consider either 1) a batch effect that has already been observed, or 2) a new, unobserved batch effect. In the context of the example this is a posterior distribution on the outcome of a new experiment when 1) an existing lab technician conducts a new experiment, or 2) a new lab technician conducts a new experiment. The difference is in either utilizing an existing batch effect posterior, or using the super population variance posterior to generate a new batch effect (Gelman *et al.*, 2003).

2.2 High-dimensional Multivariate Data

For multivariate settings a statistical model analogous to the model above may also be utilized. Observations of dimension d are specified as $\mathbf{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_{ij}$. The Bayesian hierarchy for the batches may then be given by

$$\begin{array}{lll} \boldsymbol{\mu} \mid \boldsymbol{\Sigma}_\mu \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_\mu), & \boldsymbol{\beta}_i \mid \boldsymbol{\Sigma}_\beta \overset{iid}{\sim} \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta), & \boldsymbol{\epsilon}_{ij} \mid \boldsymbol{\Sigma}_\epsilon \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon), \\ \boldsymbol{\Sigma}_\mu \sim F_\mu, & \boldsymbol{\Sigma}_\beta \sim F_\beta, & \boldsymbol{\Sigma}_\epsilon \sim F_\epsilon, \end{array}$$

where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate normal distribution. Super population variances, e.g. $\boldsymbol{\Sigma}_\beta$, are now covariance matrices and their priors, e.g. F_β, are the multivariate analog of the inverse-gamma distribution, an inverse-Wishart. This multivariate ANOVA is suitable for modest values of d, although as d increases the necessary matrix computation becomes intractable. The computations are, in general, on the order of d^3. In spatial settings, as is the case for the data examined in this paper, there are several strategies to address these challenges. For example, for some spatial processes with low range behavior, one can introduce sparsity through a tapering approach, i.e. direct multiplication of the covariance with a sparse correlation matrix, as Furrer *et al.* (2006) have done to interpolate a high-resolution spatial field of precipitation. Alternatively, to adequately capture long range dependency, Cressie and Johannesson (2008), and Stein (2008) have suggested methods for covariance matrices essentially of the form $\boldsymbol{\Sigma} = \theta_0\mathbf{U} + \theta_1\mathbf{I}$, where $\mathbf{U}$ is an unknown matrix with low rank, then use this strategy to examine total column ozone satellite measurements.

Another approach is to use covariance functions commonly seen in geostatistical settings (Handcock and Stein, 1993), in which the overarching assumption is that dependence tapers off as distance increases. This allows the super population parameter space to be reduced substantially. The exponential covariance function, a special case of the more general Matérn, is defined by a range and sill, ρ, σ^2, respectively. For a pair of locations, x_1, x_2, with distance $h = \text{dist}(x_1, x_2)$, it is $C(h) = \sigma^2 \exp[-(h/\rho)^2]$, where an appropriate distance measure must be used. Data on a sphere, for example, uses great circle distances rather than Euclidean.

Given the range, a high-dimensional multivariate dataset can be analyzed using the decomposed posterior of (1). More specifically, a super population variance conditional posterior, e.g. $\text{p}(\sigma^2_\beta \mid \{\mathbf{Y}_{ij}\}, \sigma^2_\epsilon)$, can be seen to be closely approximated by an inverse-

gamma distribution (Furrer *et al.*, 2012), thus a scalar variance component as in the univariate setting. However, since batch effects are now spatial fields, the batch's finite population variance may be examined spatially, such that $s^2_\beta(x)$ is the finite population variance at location x.

3 Analysis

3.1 Data and Statistical Model

The data utilized in the analysis are decadal averages of global climate projections that have been assembled and archived as part of the CMIP3 project (Meehl *et al.*, 2007a) in the framework of the Fourth Assessment Report (Solomon *et al.*, 2007) for the Intergovernmental Panel on Climate Change (IPCC) over nine decadal periods 2010–2020 to 2090–2100. The data further consists of output from 15 different GCMs (CCCMA-CGCM3.1, CNRM-CM3, CSIRO-MK3.0, GFDL-CM2.0, GFDL-CM2.1, GISS-MODEL-ER, INM-CM3.0, IPSL-CM4, MIROC3.2, ECHO-G, ECHAM5/MPI-OM, MRI-CGCM2.3.2A, CCSM3.0, PCM, UKMO-HadCM3 (Meehl *et al.*, 2007b)) each run under three different greenhouse gas emissions scenarios (A2, A1B, and B1 (Nakićenović and Swart, 2000)). Although other GCMs and decadal periods are available, in order to maintain a balanced design, only those for which all scenarios with distinct forcings were available have been used. The analysis is carried out on four datasets in this format, i.e. precipitation (mm/day) and temperature (degrees Celsius), and boreal summer months June, July, August (JJA) and winter months December, January, February (DJF). Each spatial field has been interpolated to a common 128×64 longitudinal–latitudinal grid for approximately 2.8 degree resolution.

Bias of and dependence among climate models are important issues as outlined and illustrated by Tebaldi and Knutti (2007), Jun *et al.* (2008), and Knutti *et al.* (2010), and references therein. Here, we nevertheless adopt the statistical assumption that has traditionally been used when working with sets of GCMs, which is to assume that they are independently drawn from a common process representative of the consensus climate characteristics. The statistical model that we use is based on the methodology of Section 2 in which the first batch $\boldsymbol{\alpha}_0$ in the model consists of 15 effects, each representing a single GCM; and the second batch $\boldsymbol{\beta}_0$ covers the three emissions scenarios. A third batch $\boldsymbol{\gamma}_0$

represents the interaction between the two. To account for climate changes over decadal periods, batches $\boldsymbol{\mu}_1$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\mu}_2$ are included. Terms $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ represent overall linear and quadratic trends, respectively. Terms $\boldsymbol{\alpha}_1$, $\boldsymbol{\beta}_1$ represent GCM and ES linear temporal effects, respectively. Preliminary analyses of the data suggested that no higher order terms were needed. Therefore, the model of a single measurement type and season dataset is

$$\begin{aligned}\mathbf{Y}_{ijt} &= \overbrace{\boldsymbol{\mu}_0 + \boldsymbol{\alpha}_{0,i} + \boldsymbol{\beta}_{0,j} + \boldsymbol{\gamma}_{0ij}}^{\text{constant}} + t(\overbrace{\boldsymbol{\mu}_1 + \boldsymbol{\alpha}_{1,i} + \boldsymbol{\beta}_{1,j}}^{\text{linear}}) + t^2(\overbrace{\quad\boldsymbol{\mu}_2\quad}^{\text{quadratic}}) \;+\; \boldsymbol{\epsilon}_{ijt} \\ &= \underbrace{\boldsymbol{\mu}_0 + t\boldsymbol{\mu}_1 + t^2\boldsymbol{\mu}_2}_{\text{trends}} + \underbrace{\boldsymbol{\alpha}_{0,i} + t\boldsymbol{\alpha}_{1,i}}_{\text{GCM batches}} + \underbrace{\boldsymbol{\beta}_{0,j} + t\boldsymbol{\beta}_{1,j}}_{\text{ES batches}} + \underbrace{\boldsymbol{\gamma}_{0ij}}_{\text{interaction}} \;+\; \boldsymbol{\epsilon}_{ijt}, \end{aligned} \tag{2}$$

where $i = 1, \ldots, 15$, $j = 1, \ldots, 3$, $t = 1, \ldots, 9$. The dimension of each spatial observation on the sphere is $d = 128 \times 64 = 8192$, which results in over 3.3 million individual data points for each of the four sets of data. Values of t are centered, $t = -4, \ldots, 4$, and t^2 is transformed to be orthogonal to other predictors in the model, i.e. $t^2 - 20/3$, such that it can be considered as a sequential addition to the model. The notation allows GCM and ES effects to be referred to as $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. A specific batch may then be referred to more specifically by name or with the notation. For example, the constant GCM effect is either GCM or $\boldsymbol{\alpha}_0$, while the linear GCM effect with respect to time is tGCM or $\boldsymbol{\alpha}_1$. An individual batch field requires an additional index, e.g. constant GCM CNRM-CM3 is $\boldsymbol{\alpha}_{0,2}$. Note that the interaction term, GCM×ES, or $\boldsymbol{\gamma}_0$, will be referred to simply as the interaction.

Exploratory studies of the dependency structure of each spatial field revealed some non-stationarity. For example, the temperature data for months of June through August of the $\boldsymbol{\alpha}$ fields suggest a larger range parameter in the Southern Hemisphere than in the Northern. The $\boldsymbol{\beta}$ fields from the same data exhibit a patchy behavior, where patches often correspond to land masses. Naturally, one could propose flexible non-stationary covariance models on the sphere, capturing some or even most of the features in a single field. Since we need a common model for all fields within one batch, we have chosen an isotropic exponential covariance function with great circle distances. As with Furrer *et al.* (2007), we opt for a plug-in estimate of the exponential covariance range parameter. These have been algorithmically estimated by averaging range estimates that have been taken over large latitudinal and longitudinal subsections of the data. Because constant

	Precipitation		Temperature	
Batch	JJA	DJF	JJA	DJF
GCM $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$	1001	1054	1421	1706
ES $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$	1098	1011	3232	1999
Interaction $(\boldsymbol{\gamma}_0)$	780	926	1600	1669

Table 1: Range parameter ρ distances (km) used in covariance parameterization.

and linear terms within the greater batch are seen to exhibit similar ranges, a single range estimate is used for both. Table 1 shows the ranges in kilometers that have been used for each dataset.

Super population variance posteriors using these spatial correlation matrices can then be sampled through numerical integration, but can also be closely approximated by an inverse-gamma distribution. Batch posteriors are then sampled on a pointwise basis. Often times, as in this analysis, overall mean terms are considered nuisance parameters and are neglected. Due to orthogonality these terms do not have any impact in the analysis of the remaining batches.

The analysis based on model (2) has been implemented with the freely available computer software R (Ihaka and Gentleman, 1996; R Development Core Team, 2011) run on a shared server with multiple 1GHz processors and 512GB memory, although parallel computing is not leveraged. The most demanding operations are one-time calculations such as the spectral decomposition of correlation matrices, which take approximately 30 minutes each, leading to a total setup time of a few hours. After setup calculations, simulation of 1000 realizations from each posterior distribution requires a trivial amount of time.

3.2 Results

The primary general conclusion evident from the analysis is that GCM is the strongest source of variation in the output across the ensemble. Further, GCM batch variability was estimated to be large for both super population variance $\sigma^2_{\alpha_0}$ as well as for spatial finite population variance $s^2_{\alpha_0}(x)$ over most spatial locations (Figures 3–6). A second general conclusion is that precipitation variability can not be clearly explained by either GCM or ES. Rather, the interaction between the two, varies more greatly than the tGCM and

both ES and tES batches. Broadly speaking, seasonal patterns can also be seen across all batches. Precipitation for JJA and DJF has concentrated variability over regions that are reflected about the equator, exhibiting the oscillatory nature of the Intertropical Convergence Zone (ITCZ). Temperature does not exhibit such symmetries, but tends to see variability concentrated over land masses for JJA and over arctic regions for DJF. Results and figures are now covered in more depth.

Figures 2–6 provide a summary of batches and variance parameters primarily by posterior averages. Figure 2 displays the least squares estimates of the trend terms, as they are considered nuisance parameters. Figures 3–6 summarize the super population standard deviation parameters and the mean of finite population standard deviations at each spatial location. It is clear that the batch with the greatest variability, for both super and finite population, is GCM. As described in Section 2, super population variances are represented by a scalar sill parameter over the whole field. Thus, credibility intervals are easy to display for the super population standard deviations. In the credibility interval plots shown, super population variance posterior parameters have been scaled to reflect a univariate case, in order to highlight the relative uncertainties between the batches. Furthermore, a log scale has been used to better display the intervals. Finite population standard deviations are, on the other hand, summarized by their pointwise posterior means since each batch field posterior realization is a multivariate spatial field. Larger values then represent greater magnitude of variability for the batch. Precipitation variability is intrinsically linked with high precipitation, as can be seen in the ITCZ in all batches of Figures 3 and 4. The seasonal effect of this can be seen as well, with greater variability concentrated north of equator in summer months, and south of equator in winter months. For example, JJA interaction (bottom left Figure 3) shows higher values to the north of the equator, while DJF interaction (bottom left Figure 4) shows higher values to the south.

Figures 7 and 8 examine the spatial variation of each batch compared to errors. Specifically, these plots display a posterior probability on whether or not a batch's finite population standard deviation at location x, e.g. $s_{\alpha_0}(x)$, is greater than the error standard deviation, σ_ϵ. As with the example of Section 2, this is given by a posterior probability but now at a spatial location, e.g. $P[s_{\alpha_0}(x) > \sigma_\epsilon]$. This probability is empirically estimated by the proportion of times that the inequality was satisfied in posterior realizations. Some

batches have been omitted as they are uninteresting, namely 0 or 1 over virtually the entire field. For precipitation (Figure 7) GCM shows the highest variability over a large proportion of the domain, but suggests low variability over polar regions. Meanwhile, tGCM shows very few regions of high variability, primarily over the ITCZ. Batches ES and tES have been not been displayed for precipitation as they alone do not suggest many regions of high variation. Alternatively, the interaction term shows high variation over the ITCZ, suggesting that a compound effect between GCM and ES is occurring. For temperature (Figure 8), batch GCM has again been omitted because it also exhibits high variability everywhere, thus leading to the overarching conclusion that GCM is the greatest source of variation. Figure 8 displays seasonal effects at polar latitudes more clearly. Summer months show greater variability over antarctic regions for batch tGCM, while winter months show greater variability for arctic regions. In winter, batch ES and the interaction also exhibit higher variability over arctic areas. The top row of Figure 8 shows that ES has higher variability over continental regions.

Figure 9 illustrates how spatial batch variances compare with one another for winter months DJF. As stated earlier, GCM exhibits much more variability than all other batches, thus it has been omitted. For precipitation, top row of Figure 9, interaction variability is greater than that of tGCM. As a broad statement, one may look from left to right, and make the general conclusion that batch variability for precipitation can be ranked as $s_{\gamma_0}(x) > s_{\alpha_1}(x) > s_{\beta_0}(x) > s_{\beta_1}(x)$, over the majority of spatial locations x. Temperature, bottom row of Figure 9, does not allow for a simple ranking. Rather it illustrates how the amount of batch variability depends on latitudinal position. Over much of the central latitudes tES varies the most.

Figures 10–17 are displayed in pairs separately for precipitation and temperature, and for DJF and JJA. These figures utilize the posterior predictive distribution, in which future observations are generated from the posterior distribution. Further, these figures display the difference of two posterior predictive distributions, the difference between the ending and starting decadal periods. This is done in order to assess the relative change between the two time points. As discussed in Section 2, there is more than one way to consider a batch in the posterior predictive distribution. In this case, a hypothetical future GCM has been generated, while the existing ES batch fields are used. This is accomplished by using the super population variance posterior to obtain a density for a

novel GCM; and by using individual ES posteriors. Figure 10 shows the posterior predictive distribution for each ES, A2, A1B, and B1, from left to right. In each row the probability at spatial location x that the relative change between starting and ending decadal periods exceeds $0.25, 0.5$, or 0.75 mm/day is shown. This corresponds to 3.6, 7.2, or 10.8 inches of precipitation per year. Figure 11, displaying 20 and 80 percent quantiles for the relative change posterior, suggests both decreases and increases in precipitation certain regions. The same pair of plots is given for precipitation DJF (Figures 12, 13). Note that Despite using a subset of available GCMs, emissions scenario A1B (middle column) of Figures 11 and 13 can be compared directly to Figure 10.9, page 767, of Solomon *et al.* (2007). The same pairs of plots are also shown for temperature JJA (Figures 14, 15), and temperature DJF (Figures 16, 17). Seasonal differences for temperature are fairly evident. Summer temperature according the scenario A1B (Figure 14, middle column) gives high probability of a 2 degree Celsius increase primarily over land masses and the Antarctic, which is supported by Figure 10.30d, page 811 of Solomon *et al.* (2007). Winter temperature for the same scenario (Figure 14, middle column) suggest an increase over Arctic regions, as in Solomon *et al.* (2007, Fig. 10.30b, p. 811).

4 Discussion

One contribution of this paper can be seen as a way to promote general analysis of variance procedures that yield contextually relevant results for multivariate data. However, for the method to be widely applicable in similar contexts, several extensions and generalizations are needed. To specifically address known dependencies between individual fields within a batch, such as the known dependency between GCMs, the batch prior and/or likelihood should include correlation of some kind, most likely empirically estimated. As an additional batch in the statistical model one may consider multiple runs done for each GCM, in which initial conditions are perturbed. This would have the affect over increasing the overall degrees of freedom in a least squares context, and generally leads to more precise inference on super population variance parameters. Covariance parameterizations should be more flexible as well. The method currently allows for arbitrary covariance matrices. Therefore, non–stationarities could be handled with, for example, latitudinal varying covariance functions. Bivariate modeling of temperature

and precipitation, that is multivariate in another domain on top of the spatial domain, would allow for relationships between these two climate characteristics to be considered jointly. For computational comparison, this analysis should also be carried out using a sparse precision matrix parameterization.

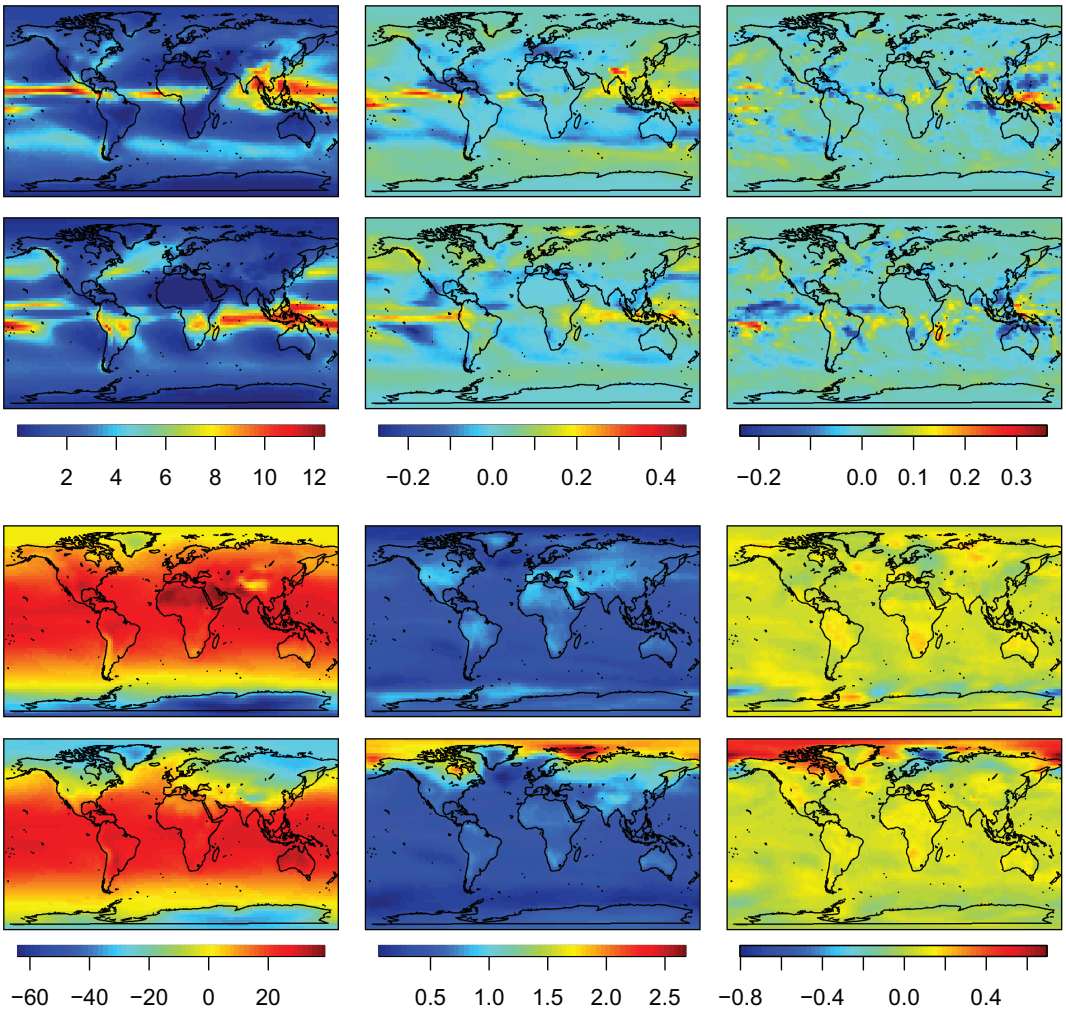

Figure 2: Trend components $\boldsymbol{\mu}_0$ (left), $\boldsymbol{\mu}_1$ (middle) and $\boldsymbol{\mu}_2$ (right) for precipitation JJA (first row), precipitation DJF (second row), temperature JJA (third row), and temperature DJF (fourth row).

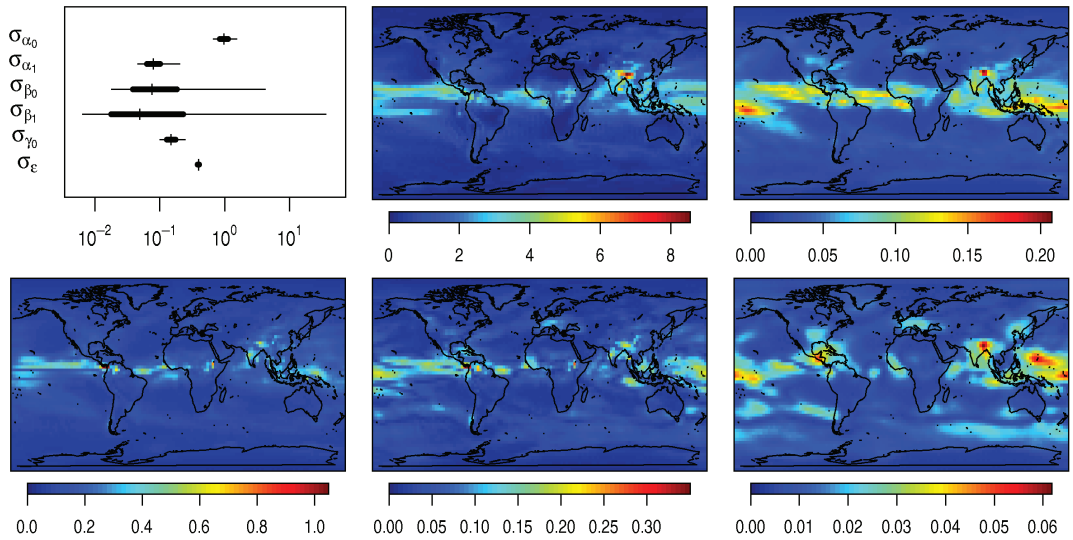

Figure 3: Precipitation JJA super population standard deviations, on log scale, credible intervals (top left). Pointwise posterior means of finite population standard deviations, on standard scale, for s_{α_0} (top middle), s_{α_1} (top right). Bottom row are posterior means of finite population standard deviations s_{γ_0} (left), s_{β_0} (middle), and s_{β_1} (right).

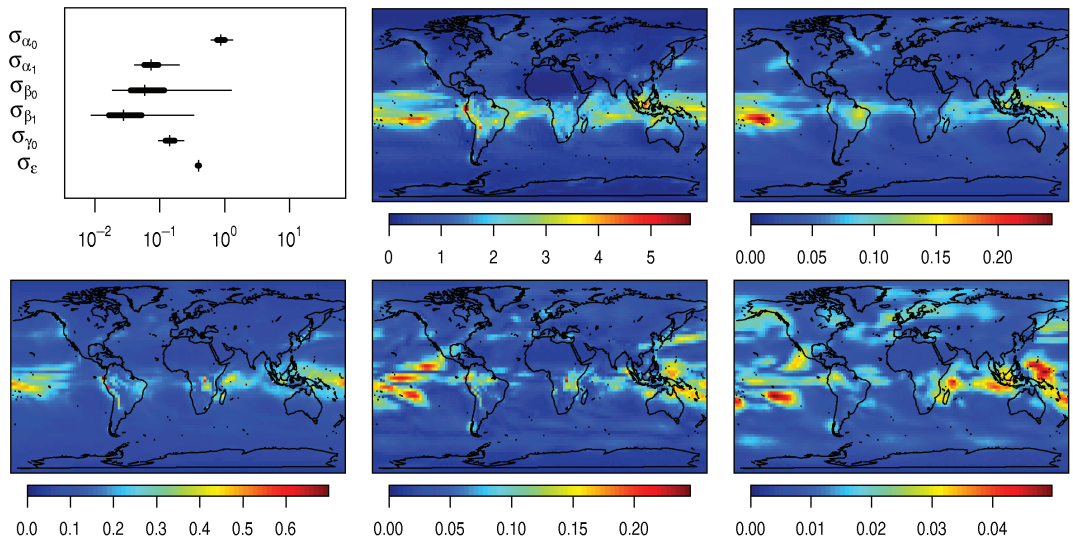

Figure 4: Precipitation DJF in same layout as Figure 3. Pointwise posterior means of finite population standard deviations, on standard scale, for s_{α_0} (top middle), s_{α_1} (top right). Bottom row are posterior means of finite population standard deviations s_{γ_0} (left), s_{β_0} (middle), and s_{β_1} (right).

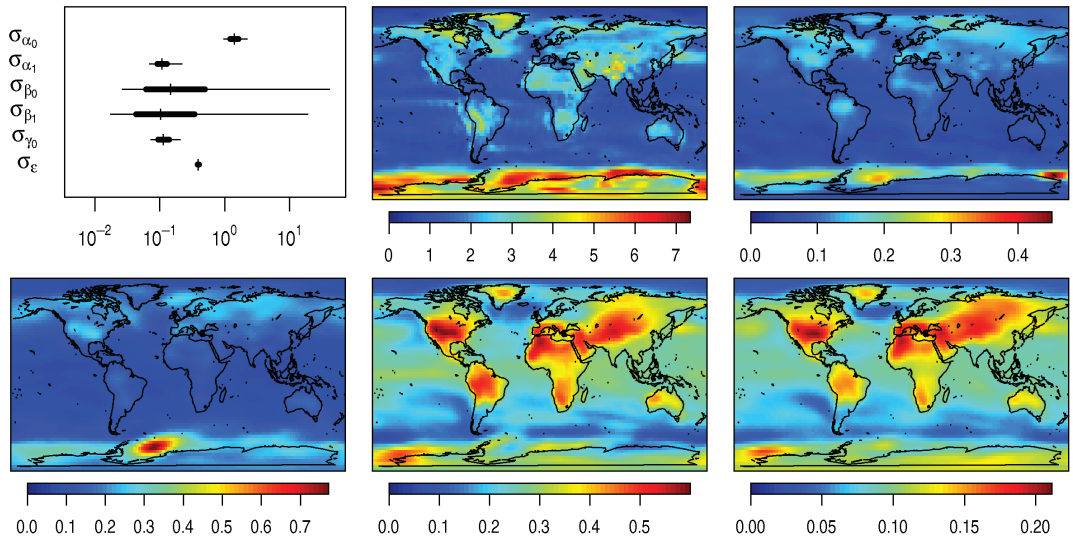

Figure 5: Temperature JJA in same layout as Figure 3. Pointwise posterior means of finite population standard deviations, on standard scale, for s_{α_0} (top middle), s_{α_1} (top right). Bottom row are posterior means of finite population standard deviations s_{γ_0} (left), s_{β_0} (middle), and s_{β_1} (right).

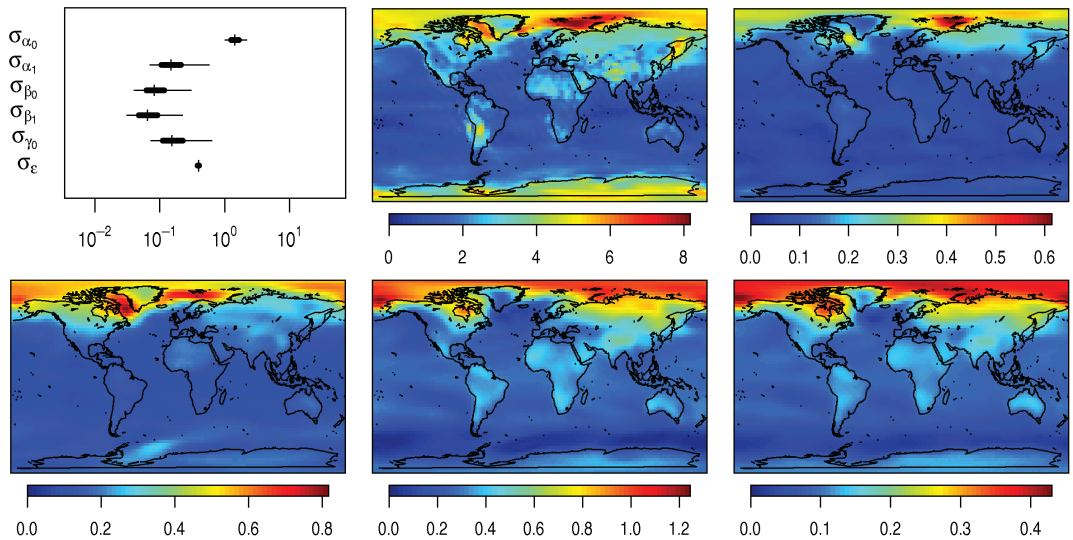

Figure 6: Temperature DJF in same layout as Figure 3. Pointwise posterior means of finite population standard deviations, on standard scale, for s_{α_0} (top middle), s_{α_1} (top right). Bottom row are posterior means of finite population standard deviations s_{γ_0} (left), s_{β_0} (middle), and s_{β_1} (right).

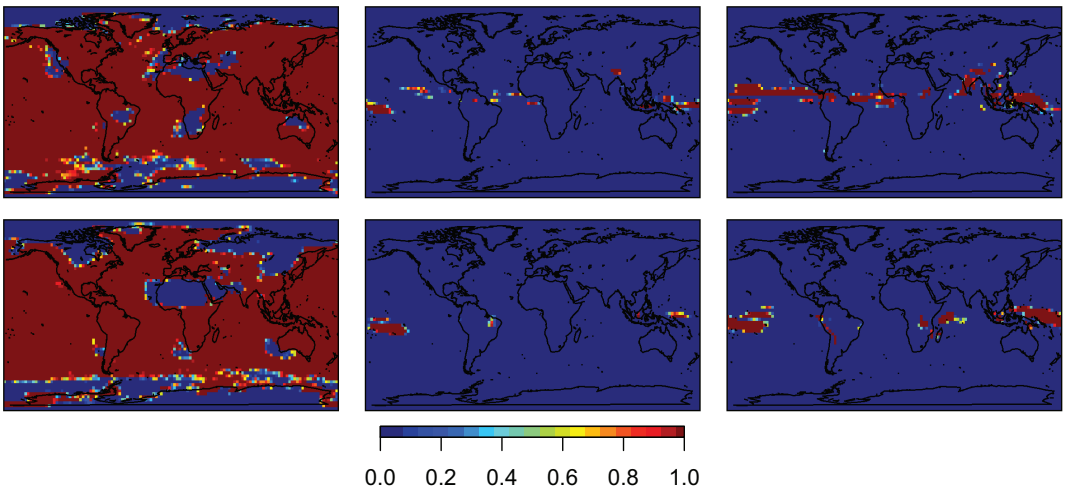

Figure 7: Precipitation JJA (top) and precipitation DJF (bottom) pointwise probabilities of $P[s_{\alpha_0}(x) > \sigma_\epsilon]$ (left), $P[s_{\alpha_1}(x) > \sigma_\epsilon]$ (middle), and $P[s_{\gamma_0}(x) > \sigma_\epsilon]$ (right).

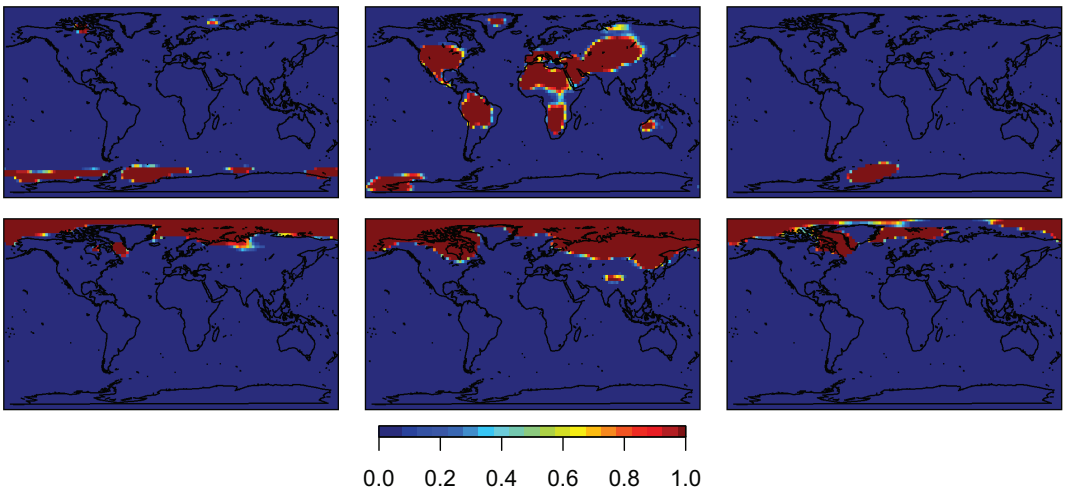

Figure 8: Temperature JJA (top) and temperature DJF (bottom) pointwise probabilities of $P[s_{\alpha_1}(x) > \sigma_\epsilon]$ (left), $P[s_{\beta_0}(x) > \sigma_\epsilon]$ (middle), and $P[s_{\gamma_0}(x) > \sigma_\epsilon]$ (right).

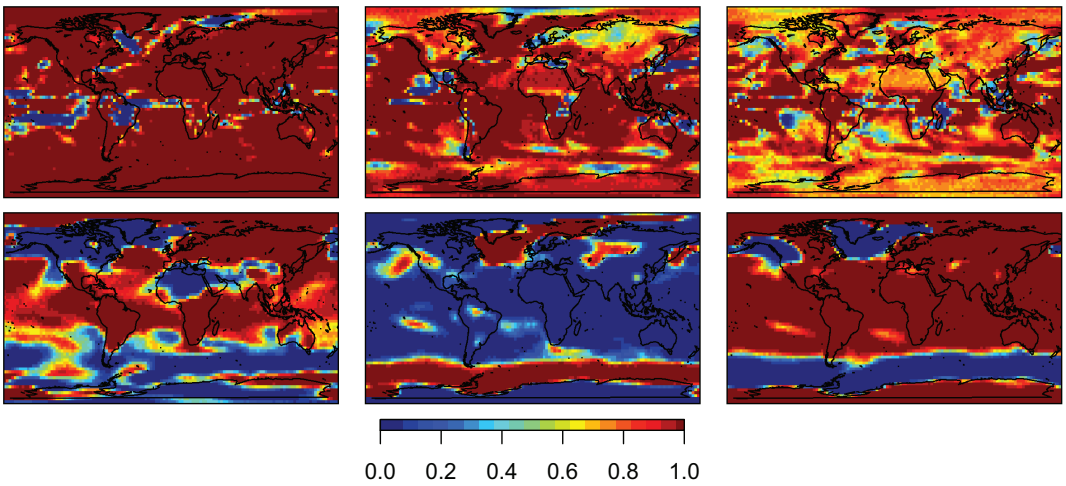

Figure 9: Relative comparison of batch variability. Pointwise probabilities of finite population standard deviations for precipitation DJF (top) are $P[s_{\alpha_1}(x) > s_{\gamma_0}(x)]$ (left), $P[s_{\alpha_1}(x) > s_{\beta_0}(x)]$ (middle), and $P[s_{\beta_0}(x) > s_{\beta_1}(x)]$ (right). Pointwise probabilities for temperature DJF (bottom) with batch comparisons $P[s_{\alpha_1}(x) > s_{\gamma_0}(x)]$ (left), $P[s_{\alpha_1}(x) > s_{\beta_1}(x)]$ (middle), and $P[s_{\beta_1}(x) > s_{\gamma_0}(x)]$ (right).

References

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B*, **70**, 209–226.

Forland, E. J., Benestad, R., Hanssen-Bauer, I., Haugen, J. E., and Skaugen, T. E. (2011). Temperature and precipitation development at Svalbard 1900–2100. *Advances in Meteorology*, **2011**, 1–14.

Furrer, R., Geinitz, S., and Sain, S. R. (2012). Assessing variance components of general circulation model output fields. *Environmetrics*, **23**, 440–450.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.*, **15**, 502–523.

Furrer, R., Sain, S. R., Nychka, D. W., and Meehl, G. A. (2007). Multivariate Bayesian analysis of atmosphere-ocean general circulation models. *Environmental and Ecological Statistics*, **14**, 249–266.

Gelman, A. (2005). Analysis of variance: Why it is more important than ever. *Annals of Statistics*, **33**, 1–31.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data*

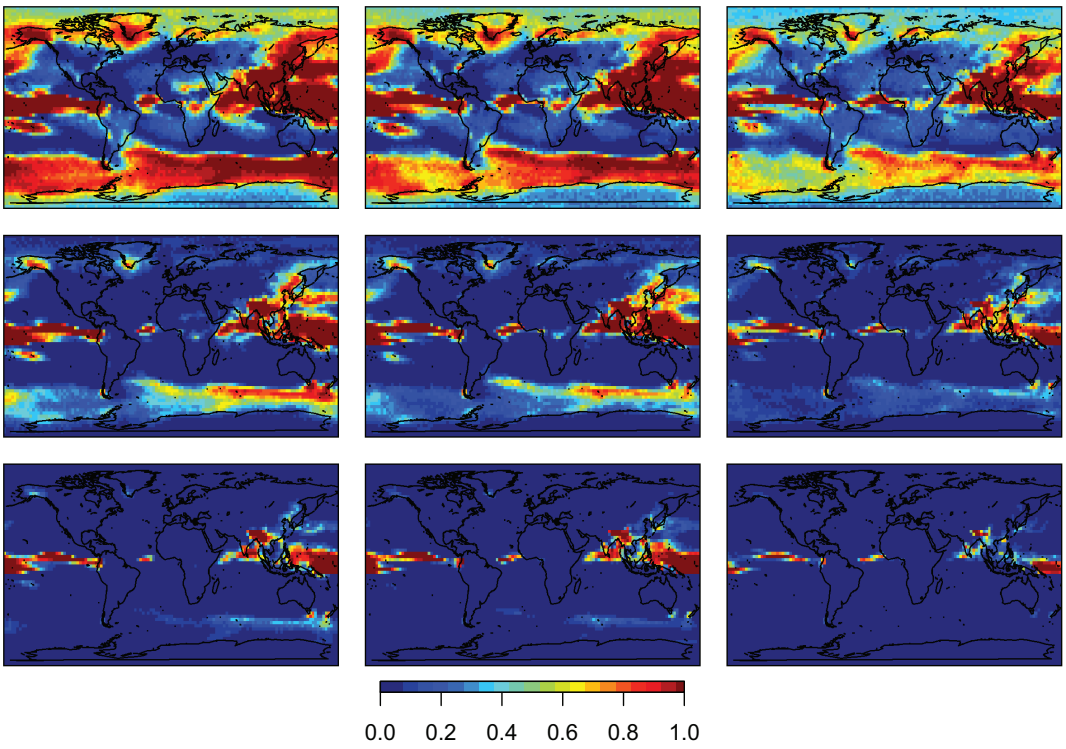

Figure 10: Precipitation JJA posterior empirical probabilities for relative increase of 0.25 (top), 0.5 (middle), and 0.75 (bottom) mm/day between decades 2010–2020 and 2090–2100. Emissions scenarios are A2 (left), A1B (middle), and B1 (right).

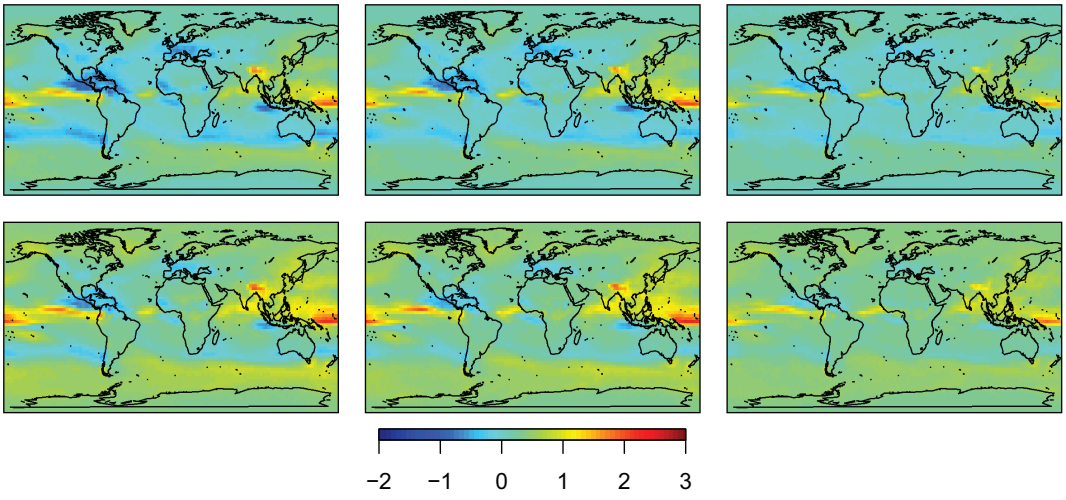

Figure 11: Precipitation JJA posterior 20% (top) and 80% (bottom) quantiles for relative change in precipitation (mm/day) between starting and ending decadal periods. Emissions scenarios are A2 (left), A1B (middle), and B1 (right).

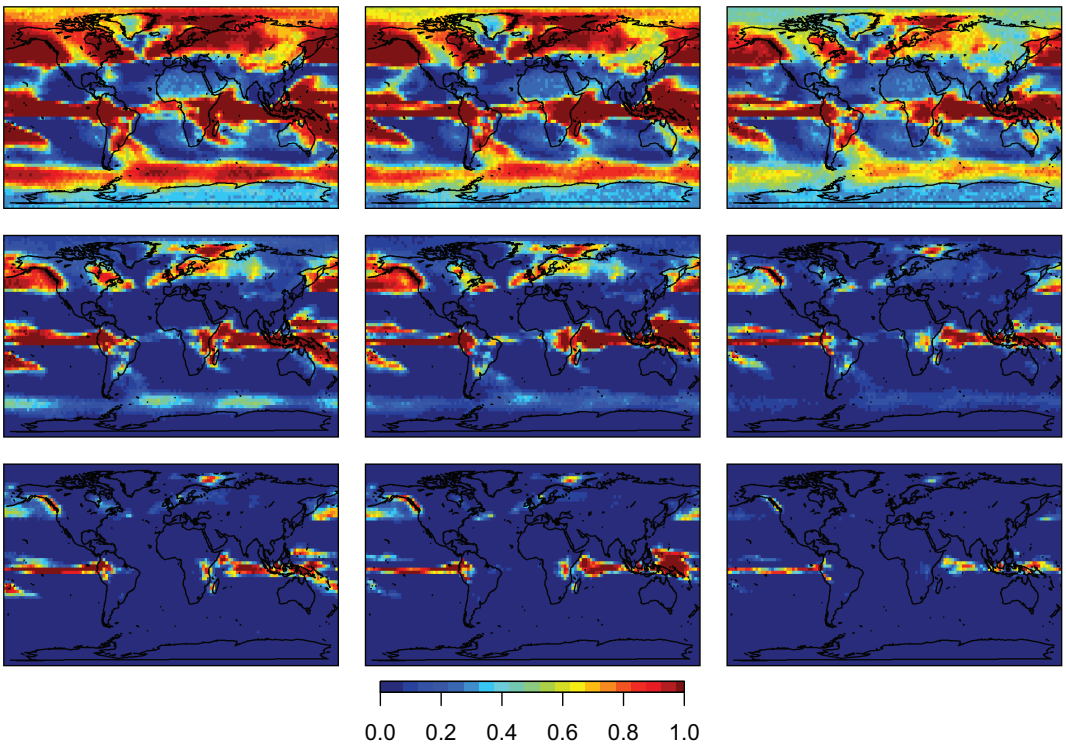

Figure 12: Precipitation DJF posterior empirical probabilities for a relative increase of 0.25 (top), 0.5 (middle), and 0.75 (bottom) mm/day between decades 2010–2020 and 2090–2100. Emissions scenarios are A2 (left), A1B (middle), and B1 (right).

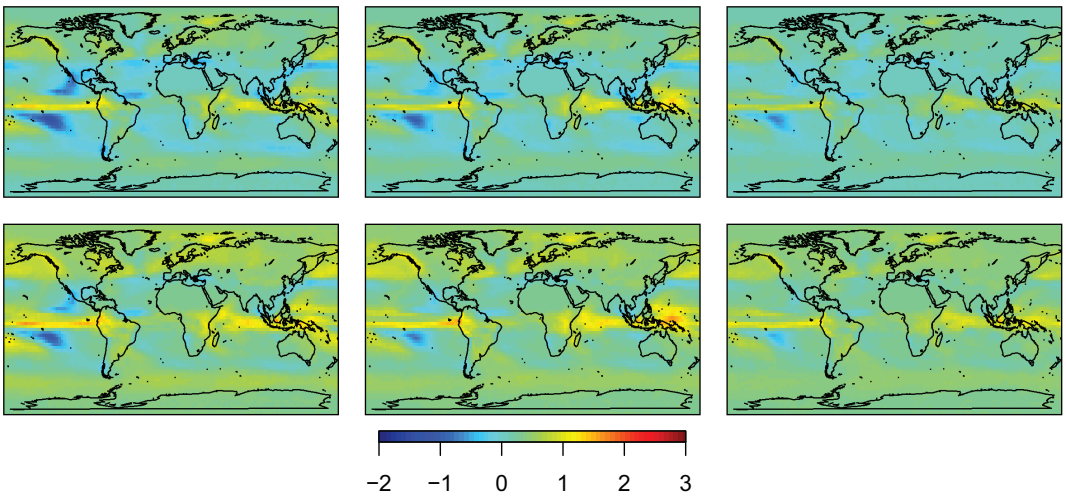

Figure 13: Precipitation DJF posterior 20% (top) and 80% (bottom) quantiles for relative change in precipitation (mm/day) between starting and ending periods (same scale as Figure 11). Emissions scenarios are A2 (left), A1B (middle), and B1 (right).

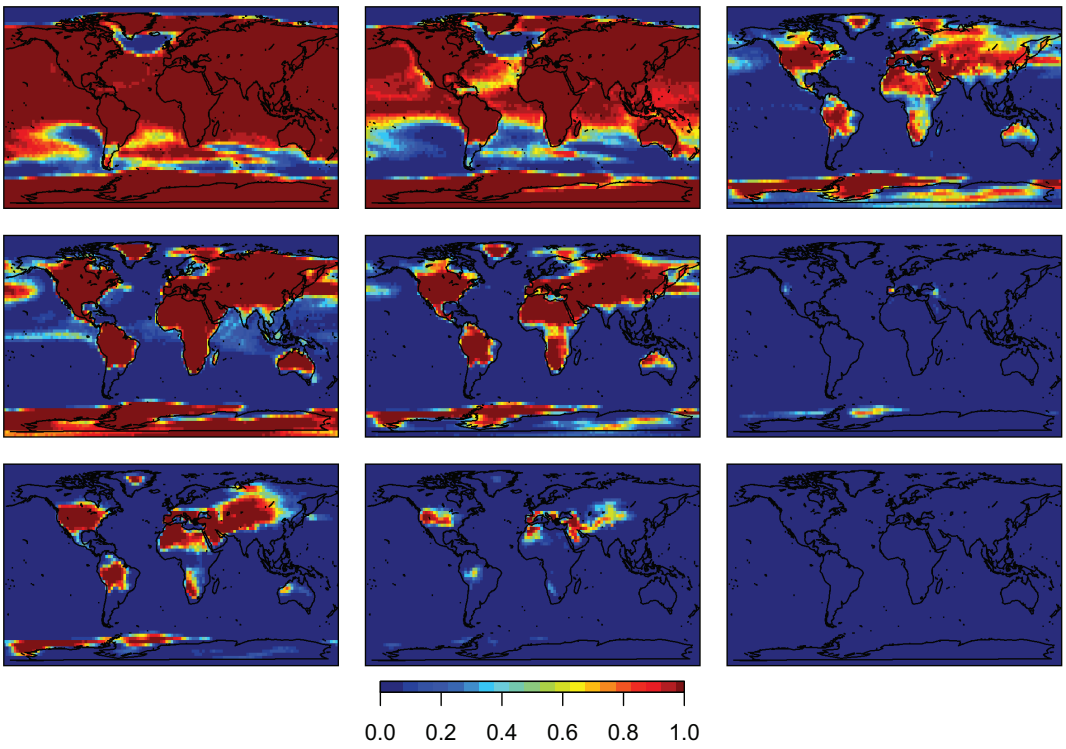

Figure 14: Temperature JJA posterior empirical probabilities for a relative increase of 2 (top), 3 (middle), and 4 (bottom) degrees Celsius between decades 2010–2020 and 2090–2100. Emissions scenarios are A2 (left), A1B (middle), and B1 (right).

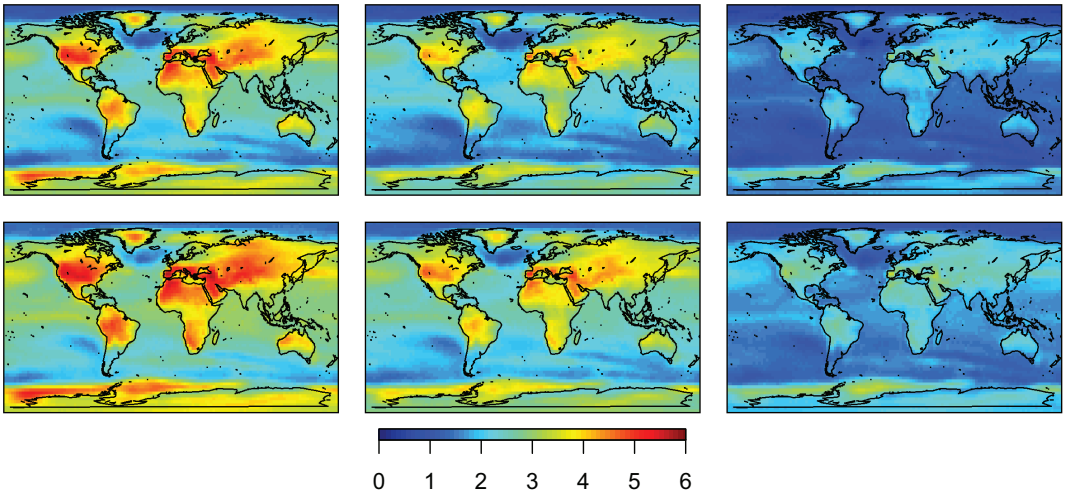

Figure 15: Temperature JJA posterior 20% (top) and 80% (bottom) quantiles for relative change in precipitation (mm/day) between starting and ending decadal periods. Emissions scenarios are A2 (left), A1B (middle), and B1 (right).

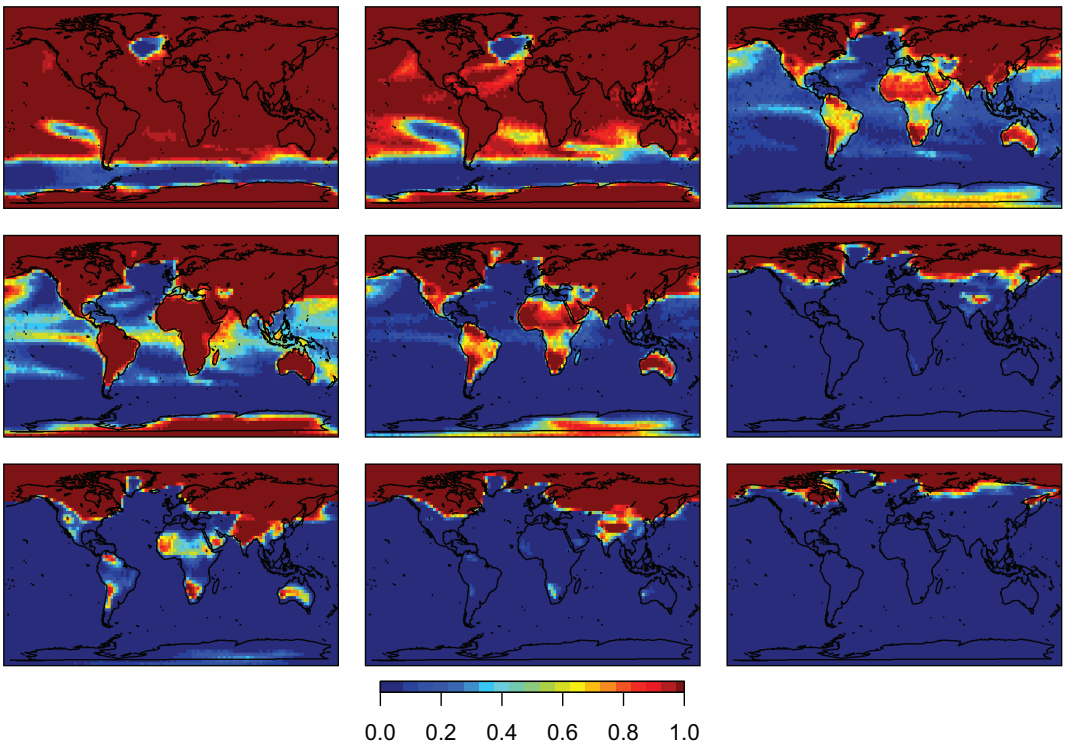

Figure 16: Temperature DJF posterior empirical probabilities for a relative increase 2 (top), 3 (middle), and 4 (bottom) degrees between decades 2010–2020 and 2090–2100. Emissions scenarios are A2 (left), A1B (middle), and B1 (right).

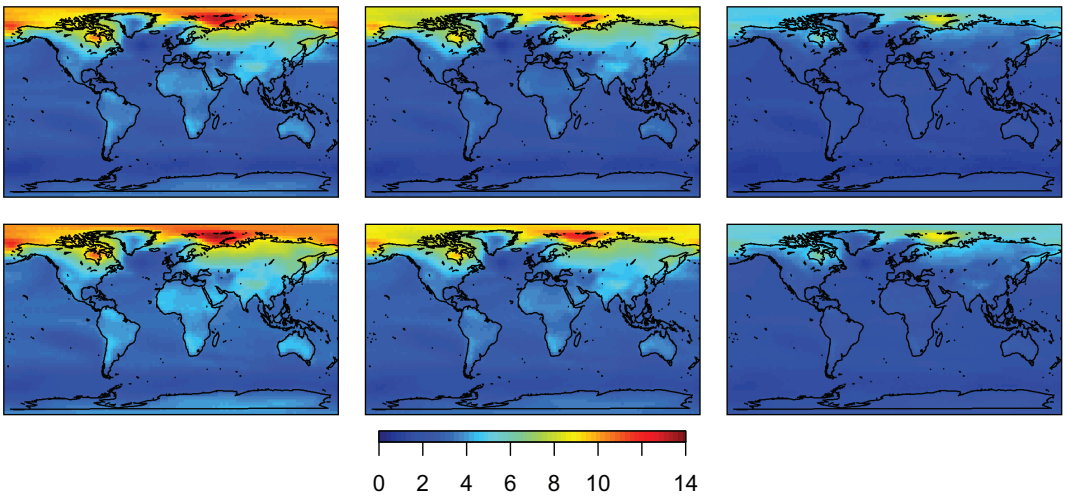

Figure 17: Temperature DJF posterior 20% (top) and 80% (bottom) quantiles for relative change in precipitation (mm/day) between starting and ending decadal periods. Emissions scenarios are A2 (left), A1B (middle), and B1 (right).

Analysis, Second Edition. Chapman & Hall/CRC.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, **35**, 403–410.

Hansen, J., Sato, M., Ruedy, R., Lo, K., Lea, D. W., and Medina-Elizade, M. (2006). Global temperature change. *Proceedings of the National Academy of Sciences*, **103**, 14288–14293.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Jin, J., Miller, N. L., and Schlegel, N. (2010). Sensitivity study of four land surface schemes in the WRF model. *Advances in Meteorology*, **2010**, 1–11.

Jun, M., Knutti, R., and Nychka, D. W. (2008). Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *Journal of the American Statistical Association*, **103**, 934–947.

Kaufman, C. G. and Sain, S. R. (2010). Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis*, **5**, 847–874.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A. (2010). Challenges in combining projections from multiple climate models. *Journal of Climate*, **23**, 2739–2758.

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., Stouffer, R. J., and Taylor, K. E. (2007a). The WCRP CMIP3 multi-model dataset: A new era in climate change research. *American Meteorological Society Bulletin*, **88**, 1383–1394.

Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, P., Gaye, A. T., Gregory, J. M., Kitoh, A., Knutti, R., Murphy, J. M., Noda, A., Raper, S. C. B., Watterson, I. G., Weaver, A. J., and Zhao, Z.-C. (2007b). Global climate projections. In Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L., editors, *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Nakićenović, N. and Swart, R., editors (2000). *IPCC Special Report on Emission Scenarios*. Cambridge University Press.

R Development Core Team (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

396 Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor,
397 M., and Miller, H. L., editors (2007). *Climate Change 2007 – The Physical Science*
398 *Basis: Working Group I Contribution to the Fourth Assessment Report of the IPCC*.
399 Cambridge: Cambridge University Press.
400 Stein, M. L. (2008). A modeling approach for large spatial datasets. *Journal of the*
401 *Korean Statistical Society*, **37**, 3–10.
402 Tebaldi, C. and Knutti, R. (2007). The use of the multi-model ensemble in probabilistic
403 climate projections. *Philosophical Transactions of the Royal Society A*, **365**, 2053–2075.

Paper IV

Assessing variance components of general circulation model output fields

Steven Geinitz & Reinhard Furrer & Stephan R. Sain

Paper published in Environmetrics, 2012.

Special Issue Paper Environmetrics

Received: 11 October 2011, Revised: 12 January 2012, Accepted: 13 January 2012, Published online in Wiley Online Library: 16 February 2012

(wileyonlinelibrary.com) DOI: 10.1002/env.2139

Assessing variance components of general circulation model output fields†

Reinhard Furrer[a]*, Steven Geinitz[a] and Stephan R. Sain[b]

Recent internationally coordinated efforts have used deterministic climate models for a common set of experiments and have produced large datasets of future climate projections. These ensembles are subject to many sources of variability, and we propose an analysis of variance procedure to quantify the contribution from several sources to the overall variation. This procedure is based on a Bayesian linear model parameterization and is applicable for large spatial data. A key feature is that individual sources of variability are modeled through batches and assessed through the batches' superpopulation variance, individual batch-level predictions, and finite population covariance. Further, for a large class of models, we show that the full posterior can be factored into conditionally independent distributions, consisting of a batch's superpopulation and batch levels. By doing so, we obviate the need for Markov chain Monte Carlo methods. Finally, this approach is applied to decadal summer temperatures for different climate models and various scenarios. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: multivariate ANOVA; spatial data; mixed model; Bayesian inference

1. INTRODUCTION

Analysis of variance (ANOVA) is often regarded as a procedure for dealing with fixed effects in a model in which various factors contribute to variation in the response. Following the strategy of Gelman (2005), we regard ANOVA as a variance component approach, considering each factor in the statistical model as random. Therefore, we use the term batch for these factors. Questions about the greater batch population are addressed with assessments of the superpopulation variance, whereas focus on the finite set of observed batch levels implies the use of the finite population variance. Geinitz *et al*. (2012) extended the Gelman (2005) approach to multivariate settings in which each batch has more levels than the degrees of freedom of the superpopulation covariance matrix. However, many problems in the environmental and biological sciences involve the analysis of highly multivariate or large spatial datasets. In such settings, it is possible to parameterize the (superpopulation) covariance matrices. We present an approach based on partitioning the posterior into conditionally independent batch posteriors, consisting of a batch's superpopulation variance and levels of the batch, which obviates the need for Markov chain Monte Carlo (MCMC) methods. We apply this method to the analysis of a climate model ensemble.

1.1. Climate models

Climate models attempt to simulate the Earth's climate system including the complex interactions among the atmosphere, land surface, ocean, sea ice, and biogeochemical processes. They are referred to as atmosphere–ocean general circulation models or simply general circulation models (GCMs). In these models, the motion of the atmosphere and the ocean is the result of discretizing the partial differential equations of fluid dynamics, constrained by external forcings (solar radiation, changes in demographics and resulting changes in land use, anthropogenic impacts on the composition of the atmosphere, etc.). However, there are many small-scale geophysical processes that cannot be resolved by the discretized model but have important feedbacks to larger scales. These processes are parameterized in the models and, because they are not derived directly from classical physics, are potential sources of model bias.

The limited spatial resolution of GCMs is often not sufficient for regional assessment of economic, social, or ecological impacts of climate change. For high(er)-resolution studies, downscaling methods can be used in which information from GCMs is used to provide a starting point for obtaining climate information on regional and even more local scales. For example, regional climate models (RCMs) use boundary conditions from GCMs to deliver high-resolution climate simulations over a bounded domain.

The development, maintenance, and use of a GCM involve extensive scientific research, software engineering, and supercomputer time. There are somewhere near 20 models that exist, mostly differing in how the unresolved (small-scale) processes are implemented and, in the numerical approaches, used to solve the partial differential equations.

* *Correspondence to: Reinhard Furrer, University of Zurich, Zurich, Germany. E-mail: furrer@math.uzh.ch*

a *University of Zurich, Zurich, Germany*

b *National Center for Atmospheric Research, Boulder, CO, U.S.A.*

† *This article is published in Environmetrics as a special issue on Advances in Statistical Methods for Climate Analysis, edited by Peter Guttorp, University of Washington, Norwegian Computing Center, Stephan R. Sain, National Center for Atmospheric Research, Christopher K. Wikle, University of Missouri.*

 Copyright © 2012 John Wiley & Sons, Ltd.

The human influence on the Earth's climate in the coming decades can only be speculated. The climate science community has henceforth defined possible (greenhouse gas) emission scenarios on which the models are based. These scenarios define different hypothetical development of the anthropogenic forcings.

The different components of a GCM generate a vast array of output. However, the most common atmospheric variables used for assessing the impacts of a changed climate are surface temperatures, precipitation, and to a lesser extent surface winds. These variables are often averaged to monthly or seasonal fields.

1.2. Statistical analysis of climate model output

Although the current generation of GCMs does not involve stochastic elements, the statistical community refers to the model output as "data" and assumes that it originates from some stochastic model. This framework is (implicitly or explicitly) required to synthesize different climate model data using a statistical framework. Early work in determining probabilities of global temperature change include those by Allen *et al*. (2000), Schneider (2001), Forest *et al*. (2002), Gregory *et al*. (2002), Knutti *et al*. (2002, 2003), and Frame *et al*. (2005). Around 2005, a shift towards fully Bayesian approaches has been made, initiated by Tebaldi *et al*. (2004, 2005), who evaluated probabilistic climate change on a regional level. These ideas have been extended, notably to a multivariate (temperature/precipitation) or temporal setting (Tebaldi and Sansò, 2007; Berliner and Kim, 2008; Buser *et al.*, 2009; Smith *et al.*, 2009) or to spatial settings (Furrer *et al.*, 2007a,b). In the same spirit, hierarchical modeling approaches have been applied to RCM by, for example, Sain and Furrer (2010), Sain *et al.* (2011a), and Christensen and Sain (2011).

It has to be stated that statistical modeling of GCM (or RCM) data is a difficult task, not only because of the large amount of data involved, the complex nonstationarities in the individual fields, and the unknown dependencies between the different climate models but also because of unsettled issues in climate science itself. More specifically, it is still a debate whether climate model output represents the "truth" (the climate) plus an error or whether climate itself is a possible realization from a set of possible climates. Naturally, the "observed" climate is inherently populated with observational error. Another fundamental issue is the choice of which models should be used in an analysis, for example, should "bad" models (in some metric) be included? Knutti *et al*. (2010) gave a detailed discussion of many of these issues (also Tebaldi and Knutti, 2007).

Statistically modeling the "mean structure" of the GCM and RCM outputs is important, but given the many uncertainties involved in climate modeling, it is to some extent more important to statistically assess and attribute these uncertainties. Two key questions are the following:

(1) What is the variability introduced through the choice of the boundary conditions compared with that of the individual modeling strategies for RCMs?
(2) What is the variability attributable to the different scenarios compared with the variability due to the individual models for GCMs?

Kaufman and Sain (2010) addressed the first of these and introduced a functional ANOVA approach for an RCM experiment (also Sain *et al.*, 2011b). Following this spirit, we develop in detail an ANOVA strategy, which Gelman (2005) outlined and which was further specified for multivariate cases by Geinitz *et al.* (2012). Further, we then analyze the output of 15 GCMs under three emission scenarios to explore their contribution to climate model output variation.

1.3. Outline

In Section 2, we briefly review a constrained Bayesian linear modeling approach towards ANOVA in multivariate contexts. Methods of modeling spatial variation and their computational demands are then reviewed, and an alternative approach for high-dimensional data is presented. In Section 3, simulation is used to illustrate the Bayesian linear modeling approach in the context of spatial statistics. Section 4 utilizes the method to identify relative contributions to overall variation from GCMs and emission scenarios.

2. BAYESIAN ANALYSIS OF VARIANCE FOR SPATIAL FIELDS

2.1. Model and notation for multivariate data

Geinitz *et al.* (2012) followed the approach of Gelman (2005) in indiscriminately considering every batch as random. When batches of the corresponding least squares can be orthogonalized and batch constraints are included, the full joint posterior distribution can be factored into independent batch posteriors, consisting of a batch's superpopulation and batch levels. The observations $\mathbf{Y}_i$ in $\mathbb{R}^d$, $i = 1, \ldots, n$, each consisting of a single batch level from B batches, written as

$$\mathbf{Y}_i = \sum_{b=0}^{B} \boldsymbol{\beta}_{j_i^b}^{(b)}, \quad i = 1, \ldots, n, \tag{1}$$

is modeled with the Bayesian multilevel model,

$$\mathbf{Y}_i \mid \left\{\boldsymbol{\beta}_{j_i^b}^{(b)}\right\}_b, \boldsymbol{\Sigma}_\epsilon \sim \mathcal{N}_d\left(\sum_{b=0}^{B-1} \boldsymbol{\beta}_{j_i^b}^{(b)}, \boldsymbol{\Sigma}_\epsilon\right), \quad i = 1, \ldots, n, \tag{2}$$

 Copyright © 2012 John Wiley & Sons, Ltd.

$$\boldsymbol{\beta}_{j^b}^{(b)} \mid \boldsymbol{\Sigma}_b \sim \mathcal{N}_d\left(\boldsymbol{\beta}_0^{(b)},\ \boldsymbol{\Sigma}_b\right), \quad j = 1, \ldots, n_b, \tag{3}$$

$$\begin{aligned} \boldsymbol{U}_b = \boldsymbol{\Sigma}_b + \frac{n_b}{n}\boldsymbol{\Sigma}_\epsilon \mid \boldsymbol{\Sigma}_\epsilon \sim W^{-1}(\boldsymbol{\Psi}_b,\ \kappa_b), \quad b = 0, \ldots, B-1, \\ \boldsymbol{\Sigma}_\epsilon \sim W^{-1}(\boldsymbol{\Psi}_\epsilon,\ \kappa_\epsilon), \end{aligned} \tag{4}$$

where $W^{-1}(\boldsymbol{\Psi}, \kappa)$ denotes an inverse-Wishart distribution with scale matrix $\boldsymbol{\Psi}$ and κ degrees of freedom. The batches $b = 0$ and $b = B$ correspond to an overall mean, or intercept, and to measurement errors, $\boldsymbol{\epsilon}_i$, respectively. Typically, zero-mean batch-level priors are assumed, that is, $\boldsymbol{\beta}_0^{(b)} = \mathbf{0}$. Setting $\boldsymbol{\Psi} = \mathbf{0}$, $\kappa = 0$ yields Jeffreys' noninformative prior $p(\mathbf{U}) \propto |\mathbf{U}|^{-(d+1)/2}$. The support of the inverse-Wishart distribution is the set of all positive definite matrices. Because $\mathbf{U}_b - (n_b/n)\boldsymbol{\Sigma}_\epsilon$ is required to be positive definite, (4) is referred to as a constrained inverse-Wishart prior. The covariance parameterization $\mathbf{U}_b = \boldsymbol{\Sigma}_b + (n_b/n)\boldsymbol{\Sigma}_\epsilon$ has previously been utilized in the context of multivariate random effects by Everson and Morris (2000).

Model specification with additional c_b constraints, for example, $\sum_j \boldsymbol{\beta}_j^{(b)} = \mathbf{0}$, combined with (2) and (3), results in improper prior distributions or intrinsic Gaussian Markov random fields (Rue and Held, 2005). With this constraint, the full joint posterior can be factored into conditionally independent batch posteriors

$$p\left(\boldsymbol{\Sigma}_\epsilon, \{\boldsymbol{\Sigma}_b, \boldsymbol{\beta}^{(b)}\}_{b=0}^{B-1} \mid \mathbf{Y}\right) = p(\boldsymbol{\Sigma}_\epsilon \mid \mathbf{Y}) \prod_{b=0}^{B-1} p(\boldsymbol{\Sigma}_b \mid \mathbf{Y}, \boldsymbol{\Sigma}_\epsilon)\ p(\boldsymbol{\beta}^{(b)} \mid \mathbf{Y}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\Sigma}_b),$$

offering straightforward marginal posteriors that can be sampled simultaneously, without using MCMC methods. The superpopulation posterior of batch B, that is, of the measurement error,

$$\boldsymbol{\Sigma}_\epsilon \mid \{\mathbf{Y}_i\} \sim W^{-1}\left(\boldsymbol{\Psi}_\epsilon + \sum_i (\mathbf{Y}_i - \widehat{\mathbf{Y}}_i)(\mathbf{Y}_i - \widehat{\mathbf{Y}}_i)^{\mathrm{T}},\ \kappa_\epsilon + n - \sum_{b=0}^{B-1} n_b\right),$$

can be found independently from other batch terms, depending on observations only through least squares residuals, with $\widehat{\cdot}$ denoting least squares estimates. The posterior of the remaining batches, $b = 0, \ldots, B-1$, is comprised of a superpopulation covariance posterior

$$\boldsymbol{\Sigma}_b + \frac{n_b}{n}\boldsymbol{\Sigma}_\epsilon \mid \{\mathbf{Y}_i\}, \boldsymbol{\Sigma}_\epsilon \sim W^{-1}\left(\boldsymbol{\Psi}_b + \sum_{j=c_b+1}^{n_b} (\widehat{\boldsymbol{\beta}}_j^{(b)} - \boldsymbol{\beta}_0^{(b)})(\widehat{\boldsymbol{\beta}}_j^{(b)} - \boldsymbol{\beta}_0^{(b)})^{\mathrm{T}},\ \kappa_b + n_b - c_b\right) \tag{5}$$

and individual batch-level posteriors

$$\boldsymbol{\beta}_j^{(b)} \mid \{\mathbf{Y}_i\}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\Sigma}_b \sim \begin{cases} \mathcal{N}_d\left(\widehat{\boldsymbol{\beta}}_j^{(b)},\ \frac{n_b}{n}\boldsymbol{\Sigma}_\epsilon\right) & j = 1, \ldots, c_b, \\ \mathcal{N}_d\left(\mathbf{P}^{-1}\mathbf{m}_j,\ \mathbf{P}^{-1}\right) & j = c_b + 1, \ldots, n_b, \end{cases} \tag{6}$$

where the distinct cases are due to added constraints and $\mathbf{P} = \boldsymbol{\Sigma}_b^{-1} + (n/n_b)\boldsymbol{\Sigma}_\epsilon^{-1}$, $\mathbf{m} = \boldsymbol{\Sigma}_b^{-1}\boldsymbol{\beta}_0^{(b)} + (n/n_b)\boldsymbol{\Sigma}_\epsilon^{-1}\widehat{\boldsymbol{\beta}}_j$. The batch-level posteriors can then be sampled from and adjusted according to constraints, to obtain posteriors for finite-population covariances, $\mathbf{S}_b$.

Replacing some or all summands of Equation (1) by $\mathbf{X}^{(b)}\boldsymbol{\beta}_{j_i^b}^{(b)}$ with $\mathbf{X}^{(b)}$, a design matrix, extends the methodology to a regression-type framework. The superpopulation covariance posteriors remain unchanged, and for the individual batch-level posteriors, the design matrices need to be added accordingly in (6).

2.2. Computational issues in superpopulation modeling

There are analytical and computational limitations of covariance estimation and hence of posterior superpopulation covariance estimation as well.

From the analytical side, the minimum number of batch levels required for (5) to be properly defined is $n_b \geq d$. However, Stein (1956) has shown that for high dimensions, d, the ratio of n_b^3/d must be small, illustrating the difficulty of covariance estimation. For moderate values of d and small difference $d - n_b$, one may consider improper Wishart distributions (Uhlig, 1994; Srivastava, 2003) for sampling. In the application of Section 4, $d - n_b$ is of the order of thousands.

Statistical procedures, such as evaluating a likelihood or sampling from a distribution, that explicitly model covariance parameters are computationally intensive for large or huge datasets. For a spatial observation of dimension d, calculating the inverse (i.e., solving a linear system) and/or determinant of the $d \times d$ covariance matrix is a computationally requisite step and is typically achieved through a factored form of the covariance matrix, such as a Cholesky decomposition. Such a direct factorization requires $O(d^3/3)$ operations and $O(d^2)$ memory. The $d(d+1)/2$ parameter space of an arbitrary covariance matrix is frequently reduced by further assumptions. Parametric covariance functions reduce the parameter space to a few dimensions, often expressed as $\boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}_b(\boldsymbol{\theta})$. Matérn covariance functions (Handcock and Stein, 1993) are commonly chosen to represent such spatial dependence, thus lending themselves well to confirmatory procedures even

 Copyright © 2012 John Wiley & Sons, Ltd.

when few observations are available. In other cases, less explicit parametric forms are used. Examples are block structures resulting from Kronecker products or sparsity assumptions. For spatial processes with primarily local behavior, one can introduce sparsity through a tapering approach (Furrer *et al.*, 2006), that is, direct multiplication of the covariance with a sparse correlation matrix. With a sufficiently sparse matrix, solving linear systems with a Cholesky factorization and calculating the determinant as a by-product can be found in times between $O(d)$ and $O(d^2)$. Rue and Tjelmeland (2002) and Lindgren *et al.* (2011), for example, looked at Gauss Markov random field approximations to Gaussian processes resulting in sparse precision matrices. To adequately capture a large-scale structure, Cressie and Johannesson (2008) and Stein (2008) have suggested exact methods for covariances essentially of the form $\mathbf{\Sigma} = \theta_0\mathbf{U} + \theta_1\mathbf{I}$, where $\mathbf{U}$ is an unknown matrix with low rank (relative to its dimension). These lead to matrix operations dictated by the value of p. In a similar category are the Gaussian process approximations (Banerjee *et al.*, 2008), implemented in our context through a regression-type model (also Sang and Huang, 2012).

However, regardless of the sparseness strategy that is taken, it cannot be maintained throughout the full sampling scheme. If sparse covariance representations exist, some demands are mitigated, but others involving the inverse (a full matrix) are still needed, for example, (5). Similarly, when sparse precision representations are available, operations involving its inverse (full) are still required. Only through explicit restriction of the parameter space, or by dictating a structural form, can sparse computation techniques be utilized.

2.3. Model extensions for spatial data

For large spatial fields, we propose parametric covariance functions $\mathbf{\Sigma}_b = \sigma_b^2\mathbf{R}(\boldsymbol{\theta}_b)$, where $\mathbf{R}$ is a correlation matrix. Following the notation of Section 2, so that likelihood is (2), and priors are

$$\boldsymbol{\beta}_{j^b}^{(b)} \mid \sigma_b^2, \boldsymbol{\theta}_b \sim \mathcal{N}_d\left(\mathbf{0},\, \sigma_b^2\mathbf{R}(\boldsymbol{\theta}_b)\right), \quad j = 1, \ldots, n_b, \tag{7}$$

$$\begin{aligned} U_b = \sigma_b^2 + \frac{n_b}{n}\sigma_\epsilon^2 \mid \sigma_\epsilon^2 &\sim \Gamma^{-1}(\psi_b,\, \kappa_b), \\ \sigma_\epsilon^2 &\sim \Gamma^{-1}(\psi_\epsilon, \kappa_\epsilon), \end{aligned} \tag{8}$$

where $\Gamma^{-1}(\psi, \kappa)$ denotes an inverse-gamma distribution with shape and scale parameters, ψ, κ, respectively. Setting $\psi = \kappa = 0$ yields the noninformative prior, $p(\sigma^2) \propto \sigma^{-2}$. The superpopulation posterior of σ_ϵ^2 is again an inverse-gamma distribution, with shape and scale parameters adjusted by least squares residual sums of squares and degrees of freedom, respectively. To obtain a more manageable form of σ_b^2, the choice of $\boldsymbol{\theta}_b$ must also be addressed. Simultaneous estimation of range and sill parameters is however a nontrivial problem. One may resort to Metropolis–Hastings steps (Metropolis *et al.*, 1953; Hastings, 1970) as employed by Kaufman and Sain (2010) or, in an empirical Bayes context (Carlin and Louis, 2000), obtain an estimate of one parameter that is then used in a Bayesian prior specification (Furrer *et al.*, 2007b). It is this latter approach that we advocate. For $\mathbf{S}_b = \sum_{j=c_b+1}^{n_b} \widehat{\boldsymbol{\beta}}_{j^b}^{(b)}\widehat{\boldsymbol{\beta}}_{j^b}^{(b)\mathrm{T}}$ and spectral decomposition, $\mathbf{R} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^{\mathrm{T}}$, with $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, the posterior of the superpopulation parameter σ_b^2 is then

$$\begin{aligned} p(\sigma_b^2 \mid \{\mathbf{Y}_i\}, \sigma_\epsilon^2) &\propto p(\sigma_b^2) \cdot \det(\sigma_b^2\mathbf{R} + \frac{n_b}{n}\sigma_\epsilon^2\mathbf{I})^{\frac{-(n_b-c_b)}{2}} \exp\left(-\frac{1}{2}\mathrm{trace}((\sigma_b^2\mathbf{R} + \frac{n_b}{n}\sigma_\epsilon^2\mathbf{I})^{-1}\mathbf{S}_b)\right) \\ &\propto (\sigma_b^2 + \frac{n_b}{n}\sigma_\epsilon^2)^{-1} \prod_{k=1}^{d} (\sigma_b^2\lambda_k + \frac{n_b}{n}\sigma_\epsilon^2)^{\frac{-(n_b-c_b)}{2}} \exp\left(-\frac{1}{2}\frac{(\mathbf{\Gamma}^{\mathrm{T}}\mathbf{S}_b\mathbf{\Gamma})_{kk}}{\sigma_b^2\lambda_k + \frac{n_b}{n}\sigma_\epsilon^2}\right), \end{aligned} \tag{9}$$

with support $\sigma_b^2 \geqslant -n_b\sigma_\epsilon^2/(n\lambda_{\max})$ and $(\cdot)_{kk}$ as the kth diagonal value. Although not a standard density, the eigenvalue transformation provides a more convenient form to sample from. A standard method, in this case inverse transform sampling, can be used once the decomposition is available.

3. ILLUSTRATION

To illustrate the method of Section 2, a simulation of a spatial process in two dimensions is carried out. We consider the model $\mathbf{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij} \in \mathbb{R}^d$ over varying dimension d, using conventional notation for batches $b = 0$ and $b = 2$. Following the notation of Section 2, batches in the model are then $\boldsymbol{\mu} = \boldsymbol{\beta}_1^{(0)}$, $\boldsymbol{\alpha}_i = \boldsymbol{\beta}_{j_i^1}^{(1)}$. For notational convenience, we will refer to the batch of interest, $\boldsymbol{\alpha}$ directly, for example, σ_α^2 rather than σ_1^2. Here, we set $n_\alpha = 10$, $n_\epsilon = 5$. The spatial process is observed on the two-dimensional regular grid $\{1, \ldots, s\} \times \{1, \ldots, s\}$, for $s = 4, \ldots, 12$, so that an observation $\mathbf{Y}_{ij}$ is a random vector of dimension $d = 16, \ldots, 144$. Batch $\boldsymbol{\alpha}$ superpopulation covariance follows a typical spatial covariance parameterization with $\mathbf{\Sigma}_\alpha = \sigma_\alpha^2\mathbf{R}(\rho)$, where ρ is a range parameter and σ_α^2 is the sill. The correlation matrix, $\mathbf{R}$, is constructed according to a Matérn covariance function with smoothness $\nu = 2.5$ and with effective range of 5, fixed over all values of d. The errors are a white noise process, $\mathbf{\Sigma}_\epsilon = \sigma_\epsilon^2\mathbf{I}$. The parameters $\sigma_\epsilon^2 = 1$ and $\boldsymbol{\mu}$ are fixed over all simulations, whereas σ_α^2 varies over 0.1, 1, or 10, to examine various magnitudes of signal-to-noise. A single realization is generated by drawing $\boldsymbol{\alpha}_i$ and $\boldsymbol{\epsilon}_{ij}, i = 1, \ldots, n_\alpha, j = 1, \ldots, n_\epsilon$, from their respective Gaussian distributions, then adding the mean $\boldsymbol{\mu}$.

We consider three different cases of model assumption with regards to ρ for Equation (7): first, the naive (and incorrect) assumption of spatial independence, $\mathbf{R} = \mathbf{I}$; second, an empirical Bayes approach with maximum likelihood estimation for ρ, $\mathbf{R} = \mathbf{R}(\widehat{\rho})$; lastly, an oracle point of view in which $\mathbf{R}$ is assumed to be known.

 Copyright © 2012 John Wiley & Sons, Ltd.

The results of a single simulation are given by uncertainty intervals of the posterior densities of σ_ϵ^2 and σ_α^2. There are 1000 draws from each; then, quantiles of the posterior sample are used for uncertainty intervals. The simulation is carried out 100 times, and the proportion of successful interval coverages and the average interval width are given. Table 1 shows that, for the naive assumption, uncertainty intervals are too narrow and have low coverage. When spatial dependence is accounted for, as with $\mathbf{R} = \mathbf{R}(\widehat{\rho})$, or $\mathbf{R}$ known, widths are adequately adjusted, and the intervals maintain coverage over all d.

4. APPLICATION

4.1. Data and detailed model specification

The methodology outlined in Section 2 is now applied to GCM output compiled and archived for as part of the CMIP3 project (Meehl *et al.*, 2007a) in the framework of the Fourth Assessment Report (Solomon *et al.*, 2007) for the Intergovernmental Panel on Climate Change.

The data consist of output from 15 different GCMs (CCCMA-CGCM3.1, CNRM-CM3, CSIRO-MK3.0, GFDL-CM2.0, GFDL-CM2.1, GISS-MODEL-ER, INM-CM3.0, IPSL-CM4, MIROC3.2, ECHO-G, ECHAM5/MPI-OM, MRI-CGCM2.3.2A, CCSM3.0, PCM, UKMO-HadCM3; see Meehl *et al.*, 2007b, for a detailed description), each run under three different greenhouse gas emission scenarios (A2, A1B, and B1, here collectively termed as SRES; Nakićenović and Swart, 2000). In this paper, we analyze nine decadal averages (2010–2020 to 2090–2100) of boreal summer months (June, July, August) temperature (°C). Each spatial field was interpolated to a common 128×64 longitudinal–latitudinal grid.

Bias of and dependence among climate models are important issues as outlined and illustrated by Tebaldi and Knutti (2007), Jun *et al.* (2008), Knutti *et al.* (2010), and references therein. Here, we nevertheless adopt the statistical assumption that has traditionally been used when working with sets of GCMs, which is to assume that they are independently drawn from a common process representative of true climate characteristics.

The first batch $\boldsymbol{\alpha_0}$ in the model consists of 15 levels, each representing a single GCM. The second batch $\boldsymbol{\beta_0}$ covers three emission scenarios. The third batch $\boldsymbol{\gamma}$ represents the interaction between the previous two. To incorporate climate change, temporal components $\boldsymbol{\mu}_1$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\mu}_2$ are included as well, with $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ representing an overall trend and $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ GCM-specific and SRES-specific temporal effects, respectively. A preliminary analysis suggested that no further terms are necessary. Hence, the model is

$$\mathbf{Y}_{ijt} = \boldsymbol{\mu}_0 + \boldsymbol{\alpha}_{0,i} + \boldsymbol{\beta}_{0,j} + \boldsymbol{\gamma}_{ij} + \boldsymbol{\mu}_1 t + \boldsymbol{\alpha}_{1,i} t + \boldsymbol{\beta}_{1,j} t + \boldsymbol{\mu}_2 t^2 + \boldsymbol{\epsilon}_{ijt},$$

where $i = 1, \ldots, n_\alpha = 15$, $j = 1, \ldots, n_\beta = 3$, $t = 1, \ldots, n_t = 9$, $n = n_\alpha n_\beta n_t = 405$, and $d = 128 \times 64 = 8192$, resulting in more than 3.3 million individual data points. Values of t are centered, $t = -4, \ldots, 4$, and t^2 is transformed to be orthogonal to other predictors in the model, that is, $t^2 - 20/3$, such that it can be considered as a sequential addition to the model. To enhance readability, we have again written the terms of Equation (1) in a more common form to eliminate additional subscripts. For example, the overall mean is $\boldsymbol{\mu}_0 = \boldsymbol{\beta}^{(0)}_{1^0_1}$, and the GCM batch levels with time covariate is $\boldsymbol{\alpha}_{1,i} = \boldsymbol{\beta}^{(5)}_{j_1^5}$ for i and j_i^5 equal to $1, \ldots, n_\alpha = 15$. The considered model is then identical to (2) containing four regression-type batches. Superpopulation variance parameters of interest are σ_b^2, with b denoting the batch.

Here, we use spatial processes on the sphere, implying that the correlation between grid points is given by $(\mathbf{R})_{rs} = \exp(-2c \sin(\vartheta_{rs}/2)/\rho)$, where ϑ_{rs} is the great circle distance between the two grid points r and s, $\rho_b > 0$ is the range, and c is the appropriate radian conversion (e.g., Furrer *et al.*, 2007b). A detailed exploratory study of the dependency structure of the individual spatial fields revealed that individual fields exhibit some nonstationarity. For example, for the $\boldsymbol{\alpha}$ fields, the range parameter seems often larger in the Southern Hemisphere than in the Northern. The $\boldsymbol{\beta}$ fields exhibit a patchy behavior, where patches often correspond to continents. Naturally, one could propose flexible nonstationary covariance models on the sphere, capturing some or even most of the features in a single field. Because we need a common model for all fields within one batch, we have chosen a simple isotropic exponential covariance function. More specifically, we have used the ranges 883, 2008, and 994 miles for $\mathbf{R}_\alpha$, $\mathbf{R}_\beta$, and $\mathbf{R}_\gamma$, respectively. Superpopulation variance posteriors (9), using these spatial correlation matrices, can then be sampled from or can be closely approximated by an inverse-gamma density. Batch-level posteriors are then sampled from on a pointwise basis using a series approximation, so that marginal means and variances of (6), each representing a single location in the field, are sampled from and adjusted according to the initial constraints. Sample variances of batch levels are then calculated to obtain finite population variance posteriors. Often times, as in this case, the overall mean term is neglected, and interest is only in batches with multiple levels.

The analysis based on the presented model has been implemented with the freely available computer software R (Ihaka and Gentleman, 1996; R Development Core Team, 2011) run on a shared server with multiple 1-GHz processors and 512GB memory, although parallelization is not utilized. The most demanding operations are one-time calculations such as the spectral decomposition of correlation matrices, which take approximately 45 min each, leading to a total setup time of a few hours. Results given in the next section are based on 1000 draws from each posterior distribution, taking only a trivial amount of time.

Markov chain Monte Carlo analysis of the data would entail considerable computation for the iterative sampling from each of the variance components and batch-level posteriors. For an arbitrary correlation matrix, each iteration would require calculations on the order of d^3, for example, here roughly equivalent to the setup time of the presented approach. Rather than utilizing the full condition posteriors that are typical of MCMC, the improper batch-level priors that have been chosen allow for conditional posteriors to be sampled from successively, for example, $p(\sigma_\epsilon^2|\mathbf{Y})$, $p(\sigma_\alpha^2|\mathbf{Y}, \sigma_\epsilon^2)$, and $p(\boldsymbol{\alpha}_i|\mathbf{Y}, \sigma_\epsilon^2, \sigma_\alpha^2)$. This decomposition does not explicitly yield computational gains but allows for full knowledge of the preceding posterior to be utilized in deriving an accurate inverse-gamma approximation and pointwise posterior approximation of the variance components and batch levels, respectively. After initial setup steps, generating a posterior batch-level sample is then performed linearly in d.

 Copyright © 2012 John Wiley & Sons, Ltd.

Table 1. Simulation results of nominal 0.95 coverage rate (average width in parenthesis) for increasing dimension d and different relative magnitudes of marginal variances.

	$\sigma_\alpha^2/\sigma_\epsilon^2 = 0.1$				$\sigma_\alpha^2/\sigma_\epsilon^2 = 1.0$				$\sigma_\alpha^2/\sigma_\epsilon^2 = 10$			
d	σ_ϵ^2	$\sigma_\alpha^2\mathbf{I}$	$\sigma_\alpha^2\mathbf{R}(\widehat{\rho})$	$\sigma_\alpha^2\mathbf{R}$	σ_ϵ^2	$\sigma_\alpha^2\mathbf{I}$	$\sigma_\alpha^2\mathbf{R}(\widehat{\rho})$	$\sigma_\alpha^2\mathbf{R}$	σ_ϵ^2	$\sigma_\alpha^2\mathbf{I}$	$\sigma_\alpha^2\mathbf{R}(\widehat{\rho})$	$\sigma_\alpha^2\mathbf{R}$
16	0.94(0.12)	0.84(0.08)	0.93(0.10)	0.94(0.10)	0.95(0.12)	0.61(0.43)	0.81(0.59)	0.96(0.56)	0.92(0.12)	0.64(3.65)	0.67(4.18)	0.94(4.12)
25	0.98(0.10)	0.81(0.06)	0.95(0.08)	0.99(0.08)	0.89(0.10)	0.58(0.33)	0.75(0.47)	0.93(0.47)	0.95(0.10)	0.69(3.00)	0.75(3.44)	0.97(3.41)
36	0.93(0.08)	0.78(0.05)	0.95(0.07)	0.97(0.07)	0.88(0.08)	0.58(0.27)	0.81(0.41)	0.88(0.40)	0.91(0.08)	0.48(2.45)	0.63(2.84)	0.93(2.85)
49	0.87(0.07)	0.77(0.04)	0.95(0.06)	0.96(0.06)	0.99(0.07)	0.58(0.23)	0.75(0.35)	0.94(0.35)	0.98(0.07)	0.63(2.12)	0.67(2.51)	0.94(2.48)
64	0.94(0.06)	0.77(0.04)	0.94(0.05)	0.96(0.05)	0.95(0.06)	0.56(0.20)	0.82(0.31)	0.96(0.32)	0.93(0.06)	0.50(1.86)	0.57(2.21)	0.93(2.20)
81	0.88(0.05)	0.78(0.03)	0.93(0.05)	0.92(0.05)	0.94(0.05)	0.57(0.18)	0.73(0.28)	0.94(0.28)	0.84(0.05)	0.65(1.68)	0.70(1.99)	0.94(1.99)
100	0.94(0.05)	0.73(0.03)	0.89(0.04)	0.93(0.04)	0.87(0.05)	0.51(0.16)	0.81(0.25)	0.95(0.26)	0.88(0.05)	0.57(1.49)	0.65(1.84)	0.97(1.84)
121	0.93(0.04)	0.82(0.03)	0.94(0.04)	0.91(0.04)	0.88(0.04)	0.57(0.15)	0.72(0.23)	0.91(0.23)	0.93(0.04)	0.57(1.37)	0.63(1.70)	0.94(1.68)
144	0.90(0.04)	0.76(0.03)	0.91(0.04)	0.96(0.04)	0.85(0.04)	0.56(0.14)	0.72(0.22)	0.93(0.21)	0.93(0.04)	0.42(1.24)	0.57(1.54)	0.93(1.55)
169	0.89(0.04)	0.70(0.02)	0.93(0.04)	0.90(0.03)	0.89(0.04)	0.61(0.13)	0.74(0.20)	0.95(0.20)	0.88(0.04)	0.44(1.15)	0.60(1.48)	0.89(1.43)
196	0.93(0.03)	0.75(0.02)	0.83(0.03)	0.91(0.03)	0.85(0.03)	0.62(0.12)	0.78(0.19)	0.95(0.19)	0.86(0.03)	0.57(1.06)	0.62(1.38)	0.91(1.37)
225	0.87(0.03)	0.76(0.02)	0.93(0.03)	0.96(0.03)	0.92(0.03)	0.58(0.11)	0.76(0.18)	0.92(0.18)	0.91(0.03)	0.53(1.00)	0.63(1.31)	0.90(1.31)

Errors are assumed to be spatially independent in all cases. Three superpopulation covariance, $\mathbf{\Sigma}_\alpha$, assumptions are compared, $\sigma_\alpha^2\mathbf{I}$, $\sigma_\alpha^2\mathbf{R}(\widehat{\rho})$, or $\sigma_\alpha^2\mathbf{R}$, respectively corresponding to white noise, estimated range, or $\mathbf{R}$ known.

 Copyright © 2012 John Wiley & Sons, Ltd.

4.2. Results

In this section, we summarize several pointwise statistics of the posterior distribution of finite population standard deviations, obtained by sampling from superpopulation (9) and then individual (6) batch-level posteriors. We first summarize finite population standard deviations, then compare them, and finally, consider posterior predictive distributions.

Figure 1 shows the batch-level estimates $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$ and the mean finite population standard deviations of batches $\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}_0$, $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}_1$, and $\boldsymbol{\beta}_1$. The top left panel does not quite represent the climate at 2050 as the quadratic correction needs to be taken into account. An orange-colored grid box in the top right panel results in a reduction of roughly 0.04 °C. The top middle panel is the average (linear) decadal adjustment. A red pixel in one of the two bottom row panels indicates a large standard deviation. Hence, the GCM data vary most over the Antarctic and the adjacent sea, whereas the scenarios differ most over the land. The mean finite population standard deviations of batches of the errors is not shown as they are of negligible interest. The superpopulation variance of the errors has first and third quartiles of 0.1537 and 0.1539, respectively. The fact that the variability in the interaction is only very locally pronounced (except for one region north of the Antarctic between the continents of Africa and South America), underpins the early uses of pattern scaling approaches to assess climate change and its variability Santer *et al.* (1990).

Given the large variability of batch $\boldsymbol{\alpha}_0$, it is natural to investigate in a next step which models (i.e., levels) contribute most to it. Figure 2 summarizes the individual GCM levels of batches $\boldsymbol{\alpha}_1$ against $\boldsymbol{\alpha}_0$. The left panel gives the root mean square posterior mean calculated over all grid boxes and area weighted. The models are nicely spread (correlation of 0.39) with no apparent outliers. The middle and right columns show posterior mean effects of the "good" model GFDL-CM2.0 and the "bad" model INM-CM3.0 (in terms of root mean square posterior mean). Within one batch, the levels have completely different features.

In the top row of Figure 3, posterior distributions are utilized to examine where precisely batches vary most in relation to the error $\boldsymbol{\epsilon}$. For each grid box, an empirical probability given by the proportion of posterior realizations in which the finite population batch standard deviations exceeds the error super population standard deviations. These probabilities can be interpreted as an indication of variation due to the batch being greater than variation due to errors. Batch $\boldsymbol{\alpha}_0$ varies strongly over nearly all locations and thus has not been displayed. Also of interest are batch finite population variance comparisons among the batches, as in the bottom row of Figure 3. The large-scale structures

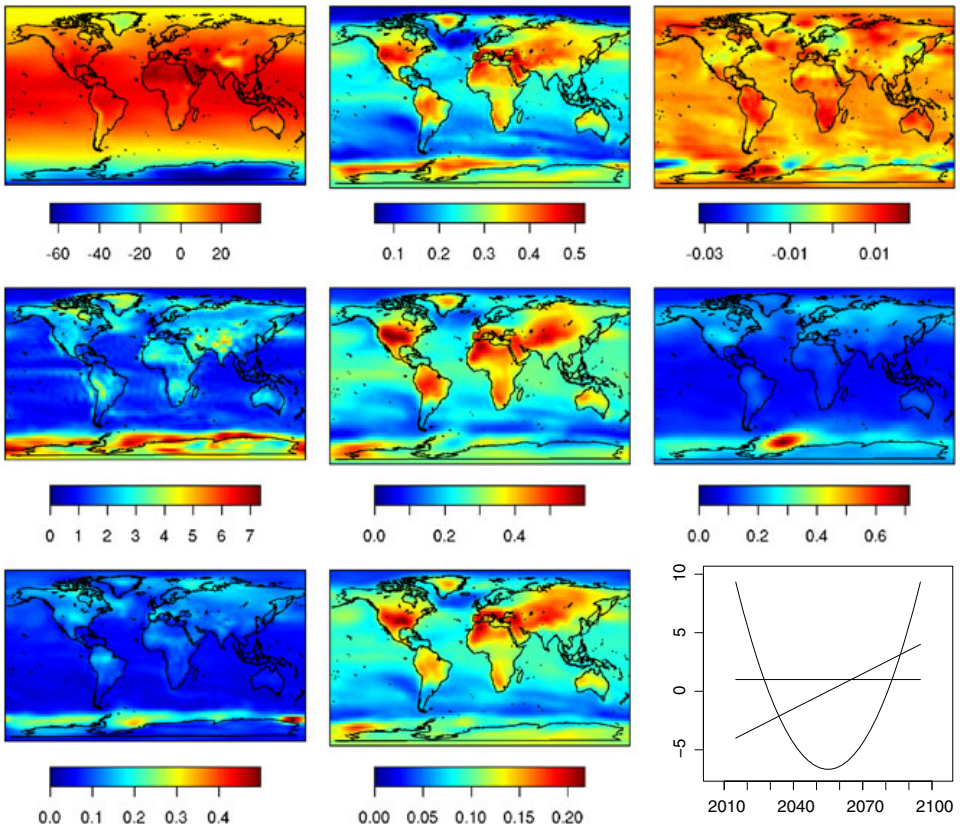

Figure 1. Batch-level estimates $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$ (top row) and mean finite population standard deviations (°C) of batches $\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}_0$, and $\boldsymbol{\gamma}$ (middle row) and $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ (bottom row). The last panel gives the temporal predictors

 Copyright © 2012 John Wiley & Sons, Ltd.

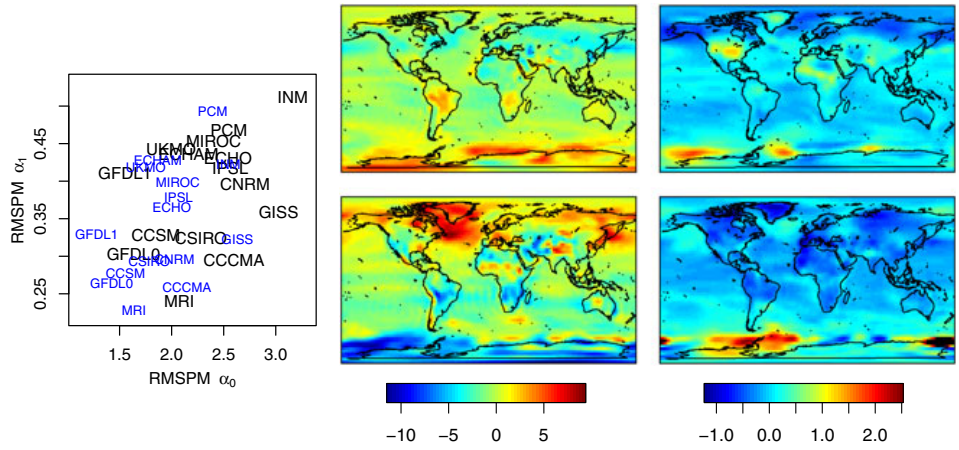

Figure 2. Root mean square posterior mean of the individual general circulation model levels of batches $\boldsymbol{\alpha}_1$ against $\boldsymbol{\alpha}_0$ (left panel), calculated over all grid boxes (black) and area weighted (blue). Posterior mean effects of batches $\boldsymbol{\alpha}_0$ (middle column) and $\boldsymbol{\alpha}_1$ (right column) for models GFDL-CM2.0 (top row) and INM-CM3.0 (bottom row). To enhance legibility, the range of batch $\boldsymbol{\alpha}_1$ is cut at 2.5. The black area in the lower-right panels contains values ranging from 2.5 to 6.92. RMSPM, root mean square posterior mean

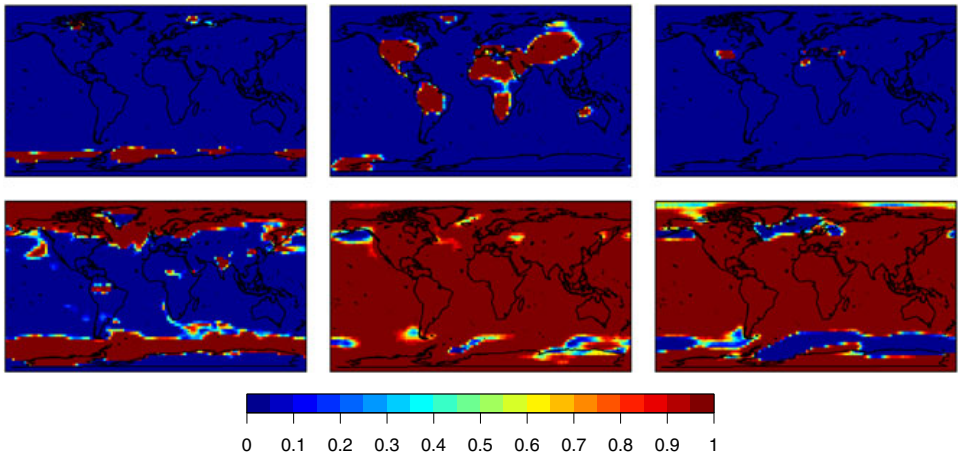

Figure 3. Pointwise empirical probabilities at location x of finite population batch standard deviations exceeding error superpopulation standard deviations, that is, $P\left(s_{\alpha_1}(x) > \sigma_\epsilon\right)$, $P\left(s_{\beta_0}(x) > \sigma_\epsilon\right)$, and $P\left(s_{\beta_1}(x) > \sigma_\epsilon\right)$ (top row) and pointwise empirical probabilities at location x comparing finite population standard deviations, that is, $P\left(s_{\alpha_1}(x) > s_{\beta_1}(x)\right)$, $P\left(s_{\beta_0}(x) > s_{\beta_1}(x)\right)$, and $P\left(s_{\beta_0}(x) > s_\gamma(x)\right)$ (bottom row)

seen in the individual panels can be guessed from Figure 1. The bottom middle panel of Figure 3 shows that the variability between the scenarios is over very large portions of the globe larger than the variability associated to the change over one century.

Posterior predictive distributions may be utilized to address the uncertainty of global climate temperature change. We illustrate this by calculating the pointwise empirical probabilities associated with a temperature increase exceeding some fixed threshold between starting and ending decadal periods. Figure 4 displays these empirical probabilities for a fictional unobserved GCM batch level and the three (observed) SRES batch levels at at least 2, 3, and 4 °C. We observe that with high probability, the specified temperature is exceeded over all land area for all scenarios and over large sea areas for the A2 and A1B scenarios. It appears that the exceedance fields are very similar between the different scenarios when the threshold temperature is adjusted by multiples of about 0.8 °C. Compared with Figure 2 in the work Furrer *et al.* (2007a), our work have a less patchy field because of the use of only one single batch level.

In a similar fashion, empirical posterior predictive quantiles can be derived. Figure 5 shows such 20% and 80% quantiles on a common scale. The top and bottom rows can be interpreted as the temperature changes that are exceeded and occur at least, respectively, with 80% probability.

 Copyright © 2012 John Wiley & Sons, Ltd.

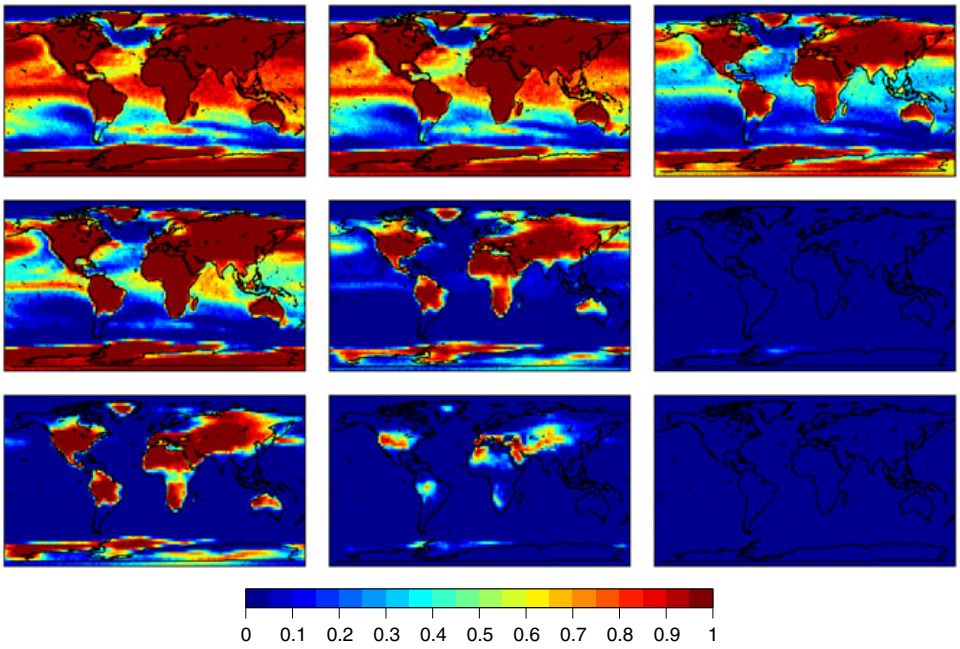

Figure 4. Pointwise empirical probabilities associated with a temperature increase of at least 2 °C (top row), 3 °C (middle row), and 4 °C (bottom row) between starting and ending decadal periods. Posterior predictive distributions are based on a fictional general circulation model batch level and on observed emission scenarios A2 (left column), A1B (middle column), and B1 (right column)

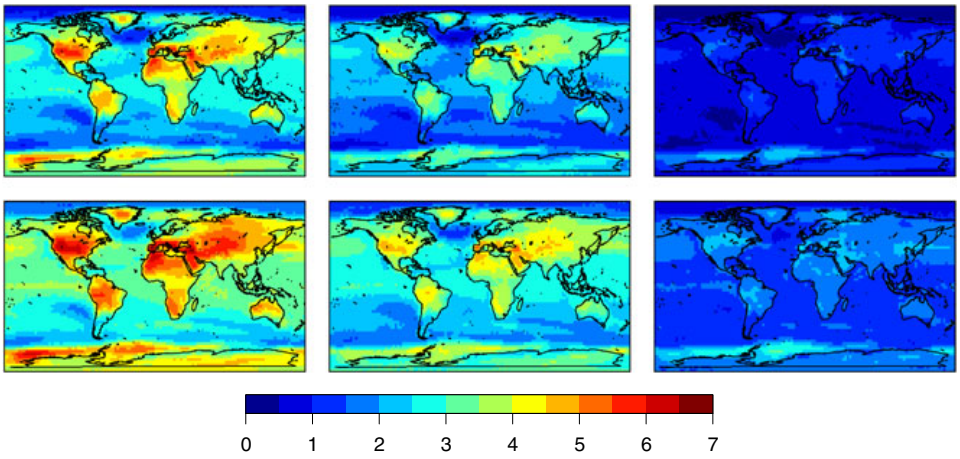

Figure 5. Pointwise empirical 20% (top row) and 80% (bottom row) quantiles (°C) between starting and ending decadal periods. Posterior predictive distributions are based on a fictional general circulation model batch level and on observed emission scenarios A2 (left column), A1B (middle column), and B1 (right column)

5. CONCLUSION

Gelman (2005) presented a paradigmatic shift in how ANOVA is approached. The work of Geinitz *et al.* (2012) and the analysis presented here extend the methodology of Gelman (2005) to multivariate and highly multivariate or spatial settings through covariance parameterization, which is a typical technique in the analysis of large spatial datasets with few or no replicates.

Deriving probability statements on random matrices, such as the posterior covariance matrix in our approach, pose not only technical challenges but also philosophical ones as well. For example, determining a batch for which the superpopulation covariance contributes more than others to overall variation requires the assumption of a criterion to compare the individual superpopulation covariance matrices. Much like Wilks' lambda (e.g., Mardia *et al.*, 1979) is defined using the determinant, other criteria could be based on other matrix characteristics,

 Copyright © 2012 John Wiley & Sons, Ltd.

such as the trace or the sum of all matrix entries. Development of such criteria and a discussion of their properties would facilitate the construction of appropriate multivariate tests in such a setting. Philosophical issues arise from the interpretation of such an analysis, as different criteria will possibly imply different conclusions.

Confirmatory statements on batch levels, or batch level means, should optimally be addressed using simultaneous inference. For example, simultaneous credible regions (e.g., Besag *et al.*, 1995 or Holmström *et al.*, 2011) provide a means to obtain uncertainty regions from batch-level posteriors. For the application of Section 4, however, the magnitude of the scalar variance components together with their uncertainty is minor in comparison to the batch-level posterior means; thus, there would not be a large distinction between simultaneous and pointwise uncertainty statements. Nevertheless, simultaneous inference procedures are necessary for further generalization of the presented method.

Independence within batch levels and between batches is a common assumption and allows an important initial step towards a more complex approach. As discussed, however, GCM batch levels are not independent. Further, a procedure that is analogous to contrasting of factor levels in traditional ANOVA would be useful in the multivariate Bayesian context as well. Given such contrasts, one may identify sets of GCMs that are similar with respect to spatial variation.

Acknowledgements

We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI), and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multimodel dataset. Support of this dataset is provided by the Office of Science, US Department of Energy.

REFERENCES

Allen MR, Stott PA, Mitchell JFB, Schnur R, Delworth TL. 2000. Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature* **407**: 617–620.

Banerjee S, Gelfand AE, Finley AO, Sang H. 2008. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B* **70**: 825–848.

Berliner LM, Kim Y. 2008. Bayesian design and analysis for superensemble-based climate forecasting. *Journal of Climate* **21**: 1891–1910.

Besag J, Green P, Higdon D, Mengersen KL. 1995. Bayesian computation and stochastic systems. *Statistical Science* **10**: 3–41.

Buser C, Künsch H, Lüthi D, Wild M, Schär C. 2009. Bayesian multi-model projection of climate: bias assumptions and interannual variability. *Climate Dynamics* **33**: 849–868.

Carlin BP, Louis TA. 2000. *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn. Chapman & Hall/CRC: London.

Christensen WF, Sain SR. 2011. Latent variable modeling for integrating output from multiple climate models. *Mathematical Geosciences*: 1–16. available online. DOI: 10.1007/s11004-011-9321-1.

Cressie N, Johannesson G. 2008. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B* **70**: 209–226.

Everson PJ, Morris CN. 2000. Simulation from Wishart distributions with eigenvalue constraints. *Journal of Computational and Graphical Statistics* **9**: 380–389.

Forest CE, Stone PH, Sokolov AP, Allen MR, Webster M. 2002. Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science* **295**: 113–117.

Frame DJ, Booth BBB, Kettleborough JA, Stainforth DA, Gregory JM, Collins M, Allen MR. 2005. Constraining climate forecasts: the role of prior assumptions. *Journal of Geophysical Research* **32**: L09702.

Furrer R, Genton MG, Nychka D. 2006. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* **15**: 502–523.

Furrer R, Knutti R, Sain SR, Nychka DW, Meehl GA. 2007a. Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. *Geophysical Research Letters* **34**: L06711.

Furrer R, Sain SR, Nychka DW, Meehl GA. 2007b. Multivariate Bayesian analysis of atmosphere–ocean general circulation models. *Environmental and Ecological Statistics* **14**: 249–266.

Geinitz S, Furrer R, Sain SR. 2012. Refining multivariate ANOVA with the Bayesian linear model. Submitted.

Gelman A. 2005. Analysis of variance: why it is more important than ever. *Annals of Statistics* **33**: 1–31.

Gregory JM, Stouffer RJ, Raper SCB, Stott PA, Rayner NA. 2002. An observationally based estimate of the climate sensitivity. *Journal of Climate* **15**: 3117–3121.

Handcock MS, Stein ML. 1993. A Bayesian analysis of kriging. *Technometrics* **35**: 403–410.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109.

Holmström L, Pasanen L, Furrer R, Sain SR. 2011. Scale space multiresolution analysis of random signals. *Computational Statistics & Data Analysis* **55**: 2840–2855.

Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**: 299–314.

Jun M, Knutti R, Nychka DW. 2008. Spatial analysis to quantify numerical model bias and dependence: how many climate models are there? *Journal of the American Statistical Association* **103**: 934–947.

Kaufman CG, Sain S. 2010. Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis* **5**: 123–150.

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA. 2010. Challenges in combining projections from multiple climate models. *Journal of Climate* **23**: 2739–2758.

Knutti R, Stocker TF, Joos F, Plattner GK. 2002. Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature* **416**: 719–723.

Knutti R, Stocker TF, Joos F, Plattner GK. 2003. Probabilistic climate change projections using neural networks. *Climate Dynamics* **21**: 257–272.

Lindgren F, Rue H, Lindström J. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B* **73**: 423–498.

Mardia KV, Kent JT, Bibby JM. 1979. *Multivariate Analysis*. Academic Press: New York.

Meehl GA, Covey C, Delworth T, Latif M, McAvaney B, Mitchell JFB, Stouffer RJ, Taylor KE. 2007a. The WCRP CMIP3 multi-model dataset: a new era in climate change research. *Bulletin of the American Meteorological Society* **88**: 1383–1394.

 Copyright © 2012 John Wiley & Sons, Ltd.

Meehl GA, Stocker TF, Collins WD, Friedlingstein P, Gaye AT, Gregory JM, Kitoh A, Knutti R, Murphy JM, Noda A, Raper SCB, Watterson IG, Weaver AJ, Zhao ZC. 2007b. Global climate projections. In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds). Cambridge UP: Cambridge.
Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**: 1087–1091.
Nakićenović N, Swart R (eds). 2000. *IPCC Special Report on Emission Scenarios*. Cambridge UP: Cambridge.
R Development Core Team. 2011. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.
Rue H, Held L. 2005. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall: London.
Rue H, Tjelmeland H. 2002. Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics* **29**: 31–49.
Sain SR, Furrer R. 2010. Combining climate model output via model correlations. *Stochastic Environmental Research and Risk Assessment* **24**: 821–829.
Sain SR, Furrer R, Cressie N. 2011a. A spatial analysis of multivariate output from regional climate models. *Annals of Applied Statistics* **5**: 150–175.
Sain SR, Nychka D, Mearns L. 2011b. Functional ANOVA and regional climate experiments: a statistical analysis of dynamic downscaling. *Environmetrics* **22**: 700–711.
Sang H, Huang JZ. 2012. A full-scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B* **74**(1): 111–132.
Santer BD, Wigley TML, Schlesinger ME, Mitchell JFB. 1990. Developing climate scenarios from equilibrium GCM results. *Technical Report No. 47*, Max Planck Institut für Meteorologie, Hamburg.
Schneider SH. 2001. What is dangerous climate change? *Nature* **411**: 17–19.
Smith RL, Tebaldi C, Nychka D, Mearns LO. 2009. Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association* **104**: 97–116.
Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds). 2007. *Climate Change 2007—The Physical Science Basis: Working Group I Contribution to the Fourth Assessment Report of the IPCC*. Cambridge UP: Cambridge.
Srivastava MS. 2003. Singular Wishart and multivariate beta distributions. *Annals of Statistics* **31**: 1537–1560.
Stein C. 1956. Some problems in multivariate analysis. *Technical Report pt. 1*, Dept. of Statistics, Stanford University and United States Office of Naval Research.
Stein ML. 2008. A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society* **37**: 3–10.
Tebaldi C, Knutti R. 2007. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A* **365**: 2053–2075.
Tebaldi C, Mearns LO, Nychka D, Smith RW. 2004. Regional probabilities of precipitation change: a Bayesian analysis of multimodel simulations. *Geophysical Research Letters* **31**: L24213.
Tebaldi C, Sansò B. 2007. Joint projections of temperature and precipitation change from multiple climate models: a hierarchical Bayesian approach. *Journal of the Royal Statistical Society: Series A* **172**: 83–106.
Tebaldi C, Smith RL, Nychka D, Mearns LO. 2005. Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles. *Journal of Climate* **18**: 1524–1540.
Uhlig H. 1994. On singular Wishart and singular multivariate beta distributions. *Annals of Statistics* **22**: 395–405.

 Copyright © 2012 John Wiley & Sons, Ltd.